Artificial Intelligence ••• (••••) •••–•••

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Collaboratively built semi-structured content and Artificial Intelligence: The story so far

Eduard Hovy [a], Roberto Navigli [b], Simone Paolo Ponzetto [b,*]

[a] *Information Sciences Institute, University of Southern California, United States*
[b] *Dipartimento di Informatica, Sapienza University of Rome, Italy*

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Knowledge acquisition
Semantic networks
Knowledge-rich methods

ABSTRACT

Recent years have seen a great deal of work that exploits collaborative, semi-structured content for Artificial Intelligence (AI) and Natural Language Processing (NLP). This special issue of the Artificial Intelligence Journal presents a variety of state-of-the-art contributions, each of which illustrates the substantial impact that work on leveraging semi-structured content is having on AI and NLP as it continuously fosters new directions of cutting-edge research. We contextualize the papers collected in this special issue by providing a detailed overview of previous work on collaborative, semi-structured resources. The survey is made up of two main logical parts: in the first part, we present the main characteristics of collaborative resources that make them attractive for AI and NLP research; in the second part, we present an overview of how these features have been exploited to tackle a variety of long-standing issues in the two fields, in particular the acquisition of large amounts of machine-readable knowledge, and its application to a wide range of tasks. The overall picture shows that not only are semi-structured resources enabling a renaissance of knowledge-rich AI techniques, but also that significant advances in high-end applications that require deep understanding capabilities can be achieved by synergistically exploiting large amounts of machine-readable structured knowledge in combination with sound statistical AI and NLP techniques.

1. Introduction

Knowledge lies at the core of Artificial Intelligence (AI) and Natural Language Processing (NLP). This fact has been continuously emphasized from the earliest pioneering work (cf. [107,123,108], to name but a few) right up to today's most recent contributions [179]. However, while substantial efforts have been devoted through the years to developing methods for the acquisition of knowledge, either manually [186,64] or automatically [54,27,143], the so-called 'knowledge acquisition bottleneck' still represents one of the main obstacles to performing complex intelligent tasks with human-level performance [97,179].

Recently, however, this stalemate has begun to loosen up. The availability of large amounts of wide-coverage semantic knowledge, and the ability to extract it using powerful statistical methods [209,53, *inter alia*], are enabling significant advances in applications requiring deep understanding capabilities, such as information retrieval [36] and question-answering engines [57]. Thus, although the well-known problems of high cost and scalability discouraged the development of knowledge-based approaches in the past, more recently the increasing availability of online collaborative knowledge

* Corresponding author.
E-mail addresses: hovy@isi.edu (E. Hovy), navigli@di.uniroma1.it (R. Navigli), ponzetto@di.uniroma1.it (S.P. Ponzetto).

0004-3702/$ – see front matter
http://dx.doi.org/10.1016/j.artint.2012.10.002

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

resources has made it possible to tackle the knowledge acquisition bottleneck by means of massive collaboration within large online communities [170,203]. This, in turn, has made these online resources one of the main driving forces behind the renaissance of knowledge-rich approaches in AI and NLP – namely, approaches that exploit large amounts of machine-readable knowledge to perform tasks requiring human intelligence. Collaboratively constructed knowledge repositories, in fact, have been used as wide-coverage sources of semi-structured information and manual annotations for a wide range of AI and NLP applications, and have consistently proved beneficial. Wikipedia[1] is a case in point of this research trend, being the largest and most popular collaborative and multilingual resource of world and linguistic knowledge that contains unstructured and (semi-)structured information.

This special issue seeks to provide a comprehensive picture of the state of the art of research aimed at exploiting semi-structured resources for AI and NLP, focusing on both the acquisition of machine-readable knowledge from these resources and the effective use of this knowledge to perform AI and NLP tasks. The initial pioneering work on using Wikipedia for research in AI was carried out around 2005–2006, and since then there have been a very large number of high-impact contributions in this area by leading research groups. In particular, proceedings from all major conferences in AI, NLP and Information Retrieval (IR) – including IJCAI, AAAI, ACL, NAACL, EMNLP, SIGIR, WWW and CIKM, to name only the top conference venues – have included in all past years a substantial number of contributions on this topic. Moreover, researchers working in this area have striven to be very active, as shown by the workshops regularly held at IJCAI, AAAI, ACL and COLING [28,29,128,66,65]. We argue that the popularity enjoyed by this line of research is a consequence of the fact that (i) it provides a viable solution to some of AI's long-lasting problems, crucially including the quest for knowledge [179]; (ii) it has wide applicability spanning many different sub-areas of AI – as shown by the papers found in this special issue, which range from computational neuroscience [154] to information retrieval [82,102], through works in knowledge acquisition [73,130,192] and a variety of NLP applications such as Named Entity Recognition [148], Named Entity disambiguation [67] and computing semantic relatedness [122,216]. Other applications include the study of modality [125,189], the generation of referring expressions [92,16], Question Answering [199], Recognizing Textual Entailment [219], text simplification [211], etc.

This introduction to the special issue is structured as follows. We begin in Section 2 by discussing the limitations of both unstructured and structured resources. We then show how semi-structured resources overcome these problems by providing a "middle ground" between structured and unstructured resources, thus bringing together the best of both worlds. In a nutshell, we argue that collaboratively-generated, semi-structured information is made up of content which is (a) semantified, (b) wide-coverage, (c) up-to-date, (d) multilingual, and (e) free in nature. Thanks to the combination of all these features, this kind of resource provides an ideal environment for acquiring large amounts of machine-readable knowledge, and deploying it effectively for a wide spectrum of AI and NLP tasks. While there are many semi-structured resources – including Wiktionary,[2] Flickr,[3] Twitter[4] and Yahoo! Answers[5] – we concentrate in this paper on Wikipedia as a case in point, seeing as it is the largest and most popular collaborative, multilingual resource of world and linguistic knowledge that contains unstructured and (semi-)structured information. We support our argument by looking at evidence from the research literature. To this end, we provide a survey of previous work centered on three main lines of research:

1. Using Wikipedia to acquire wide-coverage machine-readable knowledge (Section 3), including ontologized resources such as taxonomies and ontologies (Section 4).
2. Applying knowledge from Wikipedia to semantic Natural Language Processing tasks (Section 5).
3. Exploiting Wikipedia's distinguishing features, such as its versioned, continuously updated content (Section 6) and multilinguality (Section 7).

This way we are able not only to provide a thematic survey of work that has used collaborative, semi-structured content, but also to contextualize the various contributions collected in this special issue, introducing them into the discussion at appropriate points. We conclude with general remarks and considerations about future directions for this line of research in Section 8.

2. Why semi-structured resources?

Intelligent applications in need of knowledge can choose from many different resources, ranging from automatically generated, unstructured ones (e.g., language models computed from plain text) all the way to manually-assembled, fully-structured ones (i.e., ontologies). In the middle lie those resources which are the focus of this special issue – semi-structured ones. In this section we concentrate on the pros and cons of each kind of resource. More specifically, we try to answer the question of *what* makes semi-structured resources preferable to alternative machine-readable knowledge sources or, in other words, how they help overcome the limitations of such sources. To this end, we first look at the strengths and

[1] http://www.wikipedia.org.
[2] http://www.wiktionary.org.
[3] http://www.flickr.com.
[4] http://twitter.com.
[5] http://answers.yahoo.com.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

weaknesses of both structured and unstructured resources, and show how collaboratively-generated semi-structured content offers a powerful solution for combining the "best of both worlds" while at the same time overcoming the limitations of each.

Unstructured resources. The simplest kind of resource is just a collection of text, images or other multimedia content. Text collections (i.e., corpora) are the main kind of unstructured resource. By "unstructured" we refer to the way knowledge is formalized. In fact, while corpora provide some organizational structure (e.g., division by sentence, paragraph, section, chapter, document, etc.), they encode information only by means of strings of text. As a result the knowledge available therein is not machine-readable, as strings are readable only at the character and word level.

Resources such as raw text and unannotated corpora are easy to harvest on a very large scale, thanks to the Web. Consequently, NLP research has devoted considerable efforts in recent years to exploiting free-form running text by using the Web as a corpus [89,86,169,96,15,88,198,134, *inter alia*]. Knowledge encoded within statistical models of word co-occurrence lies at the core of language models [104,81] and vector space models [176], which are both essential components of real-world applications like Machine Translation and Information Retrieval. Moreover, work based on the so-called *distributional hypothesis* – namely, the idea that similar words appearing in similar contexts tend to have similar meaning – has shown that theoretically sound statistical approaches can produce robust models encoding different levels of linguistic information, including selectional preferences, lexical meaning and word similarity (see [197] for a survey). Distributional models, in turn, have been used for many different applications, ranging from the automatic acquisition of open-domain facts [151], relations [14] and full-fledged knowledge bases [32], to high-end tasks such as question answering [190], and document retrieval and clustering [103]. Notwithstanding these major achievements we can identify at least two fundamental limitations of using raw, unstructured data:

(a) *The knowledge gap*. Statistical models induced from very large amounts of text are known to be extremely powerful and robust [68]. However, they all still suffer from the knowledge acquisition bottleneck – i.e., they are not able to automatically acquire *all* the knowledge required for complex inference chains [49]. For instance, common sense knowledge (*birds can fly*) is almost never overtly expressed within the language data from which structured knowledge can then be extracted (e.g., on the basis of co-occurrence statistics).
(b) *Degree and quality of ontologization*. Large amounts of knowledge can be extracted from the Web with high precision in a minimally supervised fashion (cf., e.g., the Open Information Extraction framework of Banko et al. [14]). However, the output of these systems is typically not ontologized – namely, it is not included within a semantic network of unambiguously defined concepts and their semantic relations – and thus does not comply to the characteristics of a fully structured knowledge base or ontology. While recent contributions have tried to tackle this issue [163,91,143,55], questions remain as to whether the quality of their output is on a par with that of manually created resources.

Structured resources. There are many different kinds of machine-readable structured resources, depending on the level of information they encode (cf. the Ontology Learning Layer Cake presented by Buitelaar et al. [26]). These include:

- *Thesauri*. Thesauri are collections of related terms, some of which like, for instance, Roget's [172] and the Macquarie [19] Thesaurus, are specifically focused on synonyms and antonyms.
- *Taxonomies*. The next level of structuring is that of a taxonomy, which is a hierarchically-structured classification of terms (e.g., a computer scientist *is-a* person). An example of a taxonomy is provided by the Open Directory Project,[6] which categorizes Web pages into a number of domains.
- *Ontologies*. An ontology is a fully-structured knowledge model, including concepts, relations of various kinds and, possibly, rules and axioms. Concepts are the basic units of an ontology and identify meanings that belong to models of different domains. If an ontology is lexicalized, a concept is associated with one or more terms that express it by means of language. Examples of lexicalized and general-purpose ontologies are WordNet [56] and Cyc [98].

Manually-assembled repositories of structured knowledge contain information of the highest quality, since their content reflects the contribution of experts such as lexicographers and ontologists. This knowledge, in turn, has been shown to be beneficial for virtually all kinds of intelligent applications.[7] However, the acquisition and application of the knowledge found within fully-structured, manually-assembled resources exhibits non-trivial problematic issues:

(c) *Creation and maintenance effort*. As is often the case with human-labeled data, the manual creation of these resources does not scale and is extremely time-consuming, while the extracted knowledge is difficult to maintain – i.e., keep continuously updated with the latest changes. For example, it took Cyc six years to collect over one million assertions [97], nevertheless Schubert argues that the actual amount of knowledge needed for open-domain automatic processing

[6] http://www.dmoz.org.
[7] The list of research papers using, for instance, WordNet seems endless – e.g., cf. http://lit.csci.unt.edu/~wordnet/ and http://www.d.umn.edu/~tpederse/wnsim-bib.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

should be two or three orders of magnitude greater, which, given the reported progress rate, would require further decades of manual work [179].

(d) *Coverage*. Covering all topics, in order to enable open-domain intelligent processing, cannot be achieved by means of manual input from domain experts. For instance, WordNet provides fine-grained sense distinctions for many common nouns, but does not include information about named entities such as Condoleezza Rice or Theodor W. Adorno, or about specialized concepts such as cysteine or CYP2E1. Furthermore, manual resources tend to be culturally-biased – e.g., both Cyc and WordNet contain concepts for George W. Bush and Tony Blair, but no entry for, e.g., their German contemporary counterpart Gerhard Schröder.

(e) *Up-to-date information*. When moving to dynamic domains, such as processing news and other real-time information, knowledge has to be up to date. Since manually annotated resources rely on a limited input provided by domain experts, they cannot provide continuously updated information, e.g., Cyc still identifies Silvio Berlusconi as Italy's prime minister, and refers to his 2001–2006 government.

(f) *The language barrier*. To enable multilingual text processing, knowledge resources need to contain the lexicalizations of their concepts in different languages – e.g., encoding that the sense of bank as financial institute translates as banca, banco and Bank in Italian, Spanish and German, respectively. However, manual input of lexical knowledge implies that the effort needs to be repeated for each language of interest. Moreover, resource-poor languages are typically difficult to cover, due to the limited availability of resources such as lexicons, native speakers, etc.

Over the past decade, a variety of proposals – including online collaborative platforms such as MindPixel[8] and Open Mind[9] – have tried to overcome these limitations and make manual knowledge acquisition feasible by collecting input from volunteers [170]. This approach has recently been revamped through the framework of human computation proposed by von Ahn, which aims at acquiring knowledge from users by means of online games [7,8]. However, none of these efforts, to date, has succeeded in producing truly wide-coverage resources able to compete with standard manual resources.

Staking out a middle ground with semi-structured resources. Both unstructured and structured resources provide knowledge which can be successfully exploited for intelligent applications. Furthermore, from a general standpoint, their limitations are complementary – i.e., manually-assembled, fully-structured resources achieve high quality for lower coverage while, vice versa, statistical methods applied to unstructured data provide wide coverage for a lower quality. But although a promising future direction is to combine these two alternatives within a unified framework, in this paper (and special issue) we concentrate, instead, on semi-structured resources as a solution to the structured vs. unstructured knowledge dilemma. Since their inception a few years ago, these resources have proved themselves, in fact, to be of a quality on a par with manual resources [61]. High-quality content, in turn, is attained for virtually all domains – thus ensuring a wide coverage [76] – on the basis of a collaborative editing model.

In the following we focus our survey and discussion on Wikipedia, which is a prime example of this kind of resource, being, as it is, the largest and most popular collaborative semi-structured resource of world and linguistic knowledge. Although a great deal of work exists on other resources, such as Wiktionary [221,105,116], Flickr [178,167,183], Twitter [150, 149,164] and Yahoo! Answers [3,20,188], Wikipedia exemplifies all the essential features of this trend of research. Indeed, thanks to both its richness and free availability, all contributions contained in this special issue opt for it, choosing from the wide range of collaborative semi-structured resources, as their knowledge source for a variety of tasks (which we introduce in the remainder of this paper). To understand the advantages provided by semi-structured, collaborative resources like Wikipedia, we now turn to their content and editing model.

At its core, Wikipedia consists of a repository of encyclopedic entries (i.e., pages), each about a certain concept. Each page identifies a specific sense or entity of a nominal phrase: for instance, MOUSE refers to the animal sense of mouse, whereas MOUSE (COMPUTING) describes its 'pointing device' sense. Similarly, homonymous named entities are disambiguated by associating them with different pages – e.g., JOHN MCCARTHY (COMPUTER SCIENTIST) vs. JOHN MCCARTHY (LINGUIST).[10] Let us now focus on the Wikipedia page for ALAN TURING (Fig. 1), which we will use in the remainder of this paper as our primary source of examples. Similarly to other encyclopedias, the page provides definitional text from which much information can be extracted – for example, that Turing was a mathematician, that he is widely considered to be one of the fathers of AI, etc. However, what really makes Wikipedia stand out as a goldmine of knowledge is the fact that, in contrast to traditional encyclopedias, its text is partially structured. First, various relations exist between the different kinds of pages. These include, among others:

(i) *redirection pages*: for instance, TURING, TURING, A.M., A.M. TURING and CHRISTOPHER MORCOM all redirect to ALAN TURING;
(ii) *internal hyperlinks*: ALAN TURING links, among many other pages, to ARTIFICIAL INTELLIGENCE, CRYPTANALYSIS, ENIGMA MACHINE, etc.;
(iii) *interlanguage links*: these hyperlinks connect corresponding pages across wikipedias in different languages – e.g., ALAN TURING links to the Latin ALANUS MATHISON TURING and Russian Тьюринг, Алан pages;

[8] http://www.mindpixel.com.
[9] http://www.openmind.org.
[10] Accordingly, in the remainder of this paper we use the terms 'concept', 'entity', 'sense' and (Wikipedia's) 'page' or 'entry' interchangeably.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

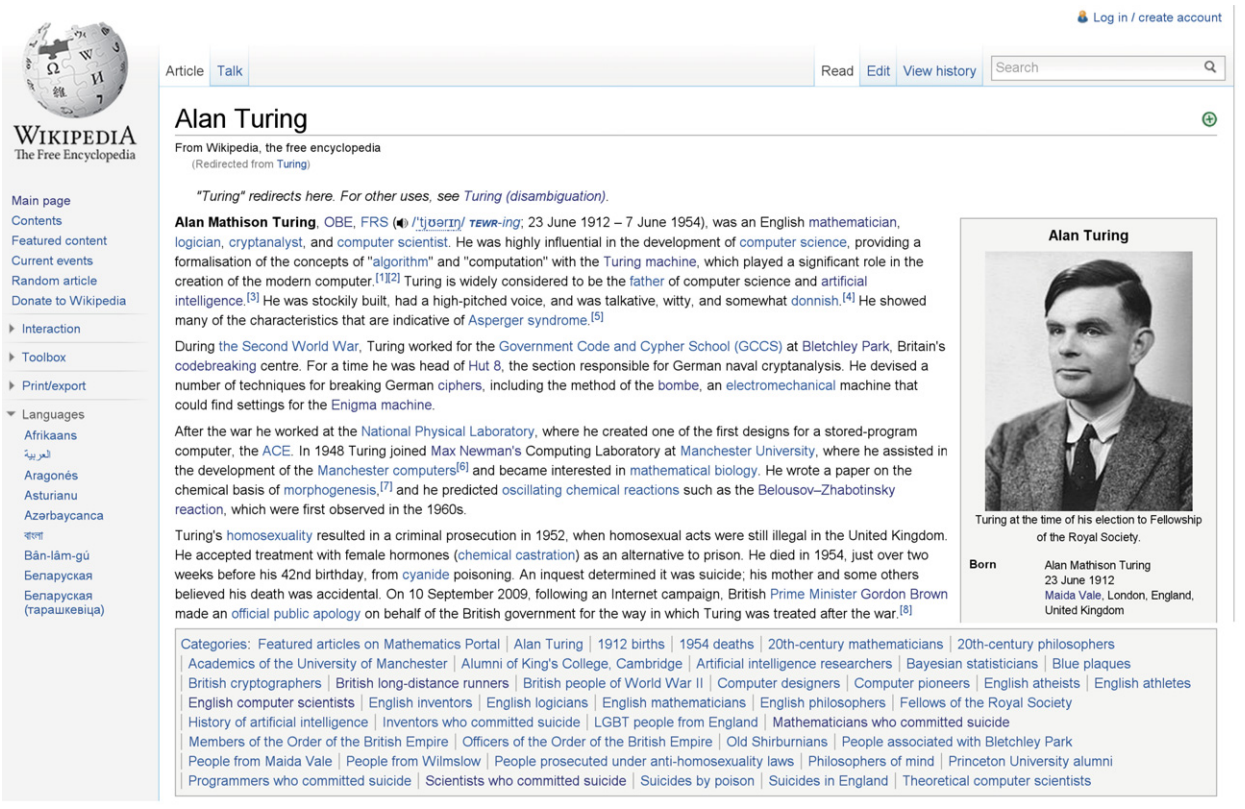

Fig. 1. Wikipedia page for ALAN TURING.

(iv) *category pages*: these are used to topically classify encyclopedic entries (e.g., ALAN TURING is categorized into MATHEMATICIANS WHO COMMITTED SUICIDE, BRITISH LONG-DISTANCE RUNNERS, etc.) and are further embedded within a hierarchical structure (e.g., MATHEMATICIANS WHO COMMITTED SUICIDE is categorized under SCIENTISTS WHO COMMITTED SUICIDE).

In addition, pages can contain infoboxes, namely tables summarizing the most important attributes of the entity referred to by a page, such as the birth and death date for a person like TURING.

On the one hand, hyperlinks and infobox tables are natural features of a web-based encyclopedia: on the other hand, the crucial aspect here is that markup annotations indirectly encode semantic content and, thus, world and linguistic knowledge manually input by human editors. For instance, categories naturally provide topic labels. Redirections, instead, are typically used to provide alternative names for a concept, that is, they model (near-)synonymy. Internal links, in turn, provide sense and entity annotations: for instance, the term ciphers within the entry for ALAN TURING is linked to CIPHER, which describes the algorithmic sense of the word (as opposed to cipher as a name for the number 0). Similarly, an occurrence of the proper name King's College (where Turing was an undergraduate student) is disambiguated, among different entities, by making it point to the entry for KING'S COLLEGE, CAMBRIDGE (as opposed, for instance, to KING'S COLLEGE LONDON). Inter-language links, instead, provide a natural way to capture sense-disambiguated translations. Finally, infoboxes encode both salient attributes for the entity of interest (e.g., the fact that Turing was British, was born in 1912 and died in 1954, etc.), as well as semantic relations (for instance, that ALONZO CHURCH was his doctoral advisor). All this content is provided by human editors on a voluntary basis – indeed, Wikipedia has been dubbed as the 'encyclopedia that anyone can edit' – and quality is achieved by massive, collaborative editing and validation.

In order to make it possible for users to provide large amounts of world and linguistic knowledge (i.e., annotated data), Wikipedia relies on a low entrance barrier: that is, (i) users are encouraged to contribute content which does not necessarily need to be structured; (ii) this content can be enhanced at any time with manual annotations, which, however, do not necessarily carry any explicit, well-defined semantics. For example, users are encouraged to hyperlink the most relevant concepts of a page, without explicitly viewing this as a sense or entity annotation process.[11] Similarly, pages and categories

[11] http://en.wikipedia.org/wiki/Wikipedia:External_links. Note that this simple, yet highly accessible, framework can later be extended with a full-fledged semantic model – cf. online collaborative platforms like Semantic MediaWiki [93], which extend Wikipedia's software platform with tools to annotate semantic data within wiki pages.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

can be categorized in order to thematically organize the encyclopedic content and, thus, provide a navigational aid. However, this thematically organized thesaurus does not necessarily consist of a semantic backbone for the concept repository: relations between Wikipedia categories are, in fact, semantically unspecified and, although methods exist for creating structured knowledge resources from their network (see Section 4.2), they do not provide *per se* a taxonomy or ontology with a full-fledged subsumption hierarchy.

Thus, the model of Wikipedia (i) relies on large amounts of manually-input knowledge, (ii) provided via massive online collaboration (iii) on the basis of semi-structured (i.e., free-form markup) content which implicitly encodes world and linguistic knowledge. We crucially argue that this model is able to alleviate, in a substantial way, the problems related to both structured and unstructured resources. This is achieved by leveraging a middle ground between structured and unstructured knowledge sources, thus taking advantage of the strengths of the resources at the two ends of the spectrum. More specifically:

(i) Being an encyclopedia, Wikipedia's content is definitional in nature. It therefore *fills the knowledge gap* by encoding large amounts of knowledge in an explicit way – e.g., the page about BIRDS mentions that 'most bird species can fly'. This in turn calls for the development of knowledge acquisition frameworks which, in contrast to mainstream approaches which typically exploit textual redundancy [50,14], concentrate instead on leveraging definitional text by means, for instance, of definition extraction techniques [142].
(ii) Wikipedia is a web of interconnected concepts and named entities, thus showing a *high degree of ontologization*. In fact, previous work has shown that its concept repository can be used 'as is' as a semantic network or taxonomy (see [160,162], for instance) or embedded within a larger knowledge base [187,22,139, *inter alia*].
(iii) In the collaborative framework upon which Wikipedia is based, *the effort required to create and maintain the knowledge repository is spread across a multitude of users*. This is to say, the high quality of its ontologized information is ensured by means of collaborative editing [61]. Each human editor, in fact, contributes his or her expertise by providing knowledge for different topics: while this anonymous crowdsourcing process is not necessarily able to produce an authoritative resource [207], it nevertheless enables scalable and open knowledge management.
(iv) Relying on vast amounts of users makes it possible to achieve a *wide coverage for virtually all domains*. In fact, thanks to its massive collaborative approach, Wikipedia is able to cover in depth not only domains pertaining to popular culture (e.g., music, movies, etc.) but also very specialized domains such as systems biology or Artificial Intelligence [76].
(v) The low entrance barrier and easy accessibility enable a *continuously updated content*, which is (i) revised to ensure high quality; (ii) kept up-to-date to reflect changes due to recent events. Content revisions, in turn, provide real-world data about how documents are collaboratively edited and revised – i.e., machine-readable records of how sentences and texts evolve through time.
(vi) The Web naturally provides a *global environment for the development of a multilingual repository of knowledge*. Wikipedia is, again, a prototype *par excellence*, since it ranks among the largest multilingual resources ever created. Currently, it covers almost 300 languages, ranging from resource-rich ones such as English and Italian to resource-poor ones like Maltese or Yiddish. A multilingual dimension is encoded not only as disambiguated translations (its so-called *inter-language links*), but also as vast amounts of text which can be exploited, for instance, to create parallel corpora [1].

Thus, semi-structured resources like Wikipedia address the problems associated with unstructured and structured resources by bringing together the best of both worlds – namely, wide coverage and high quality – by adopting a collaborative editing model (cf. Fig. 2). In the following sections we corroborate this thesis by presenting an overview of contributions from the literature which lend support to the key arguments listed above. Or, to put it another way, we now turn to a survey of contributions which highlight the advantages of using collaboratively-generated semi-structured content,[12] and explain how the papers of this special issue take up this line of research and advance it a step further.

3. Filling the knowledge gap: transforming semi-structured content into machine-readable knowledge

In this section we start looking at approaches which extract machine-readable knowledge from semi-structured content and focus on the acquisition of non-ontologized information such as terms and relations between them.

3.1. Acquiring related terms: thesaurus extraction

Thesauri consist of collections of related terms, which are grouped on the basis of their senses (e.g., based on the lexical relations of synonymy and antonymy). Sense information can be straightforwardly obtained from Wikipedia in the form of both a sense inventory (provided by its set of articles) and sense-annotated data (given by its internal links pointing to concepts and entities with unambiguous labels): consequently, Wikipedia naturally provides a source of data for automatically inducing thesauri.

[12] For a complementary survey the reader is referred to [112].

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

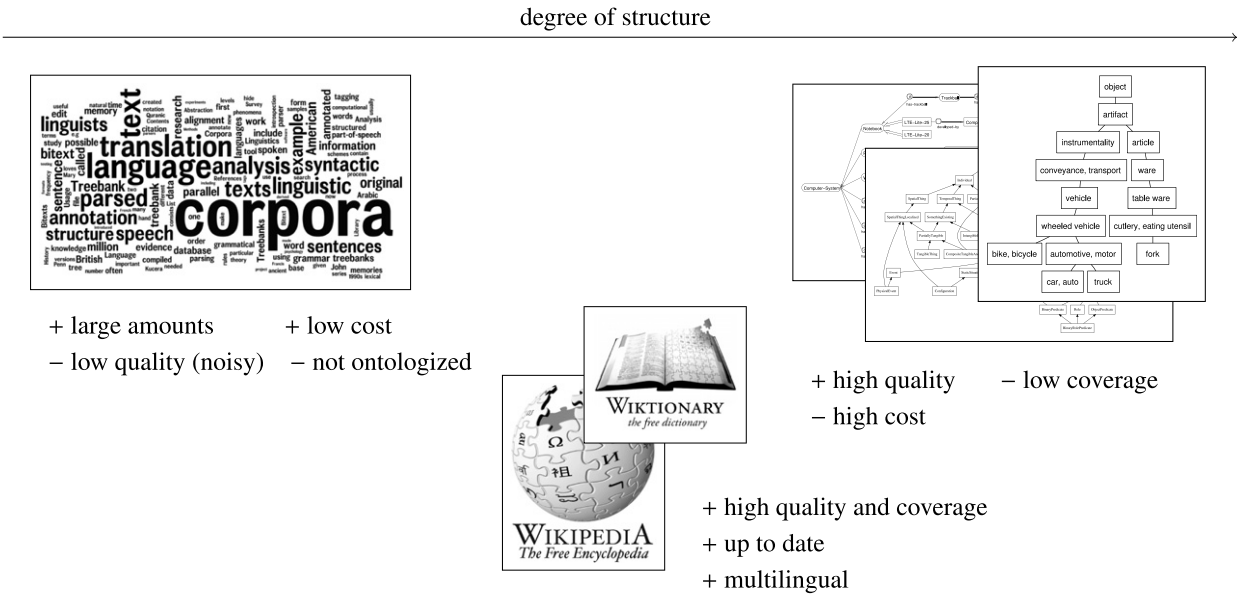

Fig. 2. Semi-structured resources such as Wikipedia stake out a middle ground between totally unstructured and fully-structured resources, thus bringing together the best of both worlds on the basis of a collaborative editing model.

Nakayama et al. [127] proposed a variety of approaches to automatically construct thesauri from Wikipedia by associating each sense (i.e., a Wikipedia page) with its most closely related senses. The simplest method is based on previous work by Schütze and Pedersen [181] and combines (i) a cosine-based scoring function and (ii) a vector space representation of documents, with (iii) a variant of the widely-used *tf*idf* measure defined using Wikipedia's hyperlinks. This computes the association strength between two pages by multiplying the number of times a specific internal link occurs within a page (*tf*) with the log-scaled inverse number of documents containing that same link (*idf*). For instance, the relatedness between Alan Turing and Artificial Intelligence is computed by counting the number of times the latter occurs as a link within the former, multiplied by the inverse number of Wikipedia pages which link to Artificial Intelligence.

A more complex association measure, named *pf*ibf*, instead, takes into account the entire structure of the graph obtained from Wikipedia's internal links. To compute the association strength between articles, this scoring function combines (i) the path frequency (*pf*), namely the number of paths between articles (each weighted by its length), with (ii) an inverse backward link frequency factor (*ibf*), which quantifies the specificity of a sense in terms of the number of backward links to its corresponding article. However, the high accuracy of *pf*ibf* comes at the cost of a high computational load, since it requires the computation of all paths between all pairs of articles in a very large graph. To resolve this, the same authors later presented a more efficient approach [79] which combines *tf*idf* vectors with second-order co-occurrence vectors [180] computed over Wikipedia's hyperlinks, and achieves a performance comparable with *pf*ibf* for a much lower computational cost.

3.2. Relation extraction

We now look at approaches which are aimed at establishing relations between entities in text – e.g., Alan Turing *born-in* London, *worked-at* Bletchley Park, etc. As in the case of thesaurus acquisition, we find that semi-structured content from Wikipedia provides large amounts of high-quality data for this task, in terms of both plain text and annotated linguistic data. A first approach, proposed by Ruiz-Casado et al. [174], collects relations between entities by analyzing the definitional sentences found in the Simple English Wikipedia.[13] From these sentences, contexts containing hyperlinks are first collected and generalized on the basis of edit distance measure and part-of-speech constraints; then, in a second step, these generalized patterns are used to harvest novel relations. Based on an automatic mapping between WordNet and Wikipedia (see Section 4.1), these relations are used to enrich WordNet with novel semantically typed *is-a* and *part-of* links between synsets.

An alternative approach to relation extraction which leverages Wikipedia's infoboxes was presented by Wu and Weld [212], who proposed to 'autonomously semantify' Wikipedia by creating new infoboxes and completing others with no explicit supervision. To this end, their system, named Kylin, leverages the shallow structure contained within Wikipedia to automatically induce supervised information extractors. Given a Wikipedia page, a document preprocessor parses the page's

[13] http://simple.wikipedia.org.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

infobox, selects relevant attributes and their value (e.g., *born-in* and MAIDA VALE), and extracts from the page sentences containing them – for instance, '[…] returned to Maida Vale, London, where Alan Turing was born on 23 June 1912 […]'. These sentences, in turn, provide automatically labeled examples for building supervised classification models which decide (i) whether an unseen sentence contains a candidate attribute-value pair, and (ii) which attribute is to be extracted. Then, given a new Wikipedia page, these classifiers are used to identify sentences which potentially contain attributes, and to extract specific attribute-value pairs from them. The result is a self-supervised system which gathers training examples and automatically learns a set of relation extractors. Self-supervision is achieved by synergistically exploiting two of Wikipedia's main characteristics – namely, structured annotations in the form of table-formatted data (i.e., infoboxes) and free-form encyclopedic text.

Matching structured annotations from infoboxes with free-form text to automatically generate labeled relational data was subsequently developed in a variety of different directions. Wu and Weld [213] learn relation-independent extractors, and combine them within TextRunner [14], in order to boost its performance. The results show that an open information extraction system can benefit greatly from a general-purpose, self-supervised extractor using dependency parse-level information (thus supporting related findings from Nguyen et al. [145]). Another extension is to enrich the feature set of Kylin's supervised relation extractors by means of Web-harvested lexicons [75], thus generalizing over lexical features. Finally, this self-supervised framework can be applied to the task of matching infoboxes from wikipedias in different languages, in order to perform cross-lingual resource mapping and integration [2].

Take-home message 1. Although their output is relatively low-level from a semantic standpoint, methods for thesaurus and relation extraction introduce us to a *leitmotif* – namely, *generating semantics by exploiting the shallow structure found in Wikipedia* – which we will encounter frequently in the remainder of this article. This simple, yet powerful idea is, in fact, common to the majority of approaches that are going to be reviewed in the following sections. Its applicability is made possible by the high quality and quantity of the collaboratively-generated, manually-curated input, which requires less noise tolerance compared to, for instance, processing Web text.

4. Degrees of ontologization: building and enriching ontologies from semi-structured content

In this section we look at resources which are created from collaboratively-generated, semi-structured content and, in contrast to thesauri and repositories of relational knowledge, provide truly semantic information, i.e., they encode semantic relations between *disambiguated concepts* and *entities*. These include taxonomies, as well as their generalization to fully-structured knowledge models, i.e., ontologies. Exploiting semi-structured content to build ontologized resources is a well-explored line of research, and is taken a step further by the two papers on YAGO2 [73] and WikiNet [130] included in this special issue, as well as by recent work on BabelNet [139], among others. We accordingly devote a detailed discussion to it. First, in Section 4.1, we look at approaches integrating semi-structured with structured information. Next, in Section 4.2, we consider taxonomic and ontological resources built by transforming semi-structured content into fully structured knowledge bases.

4.1. Integrating semi-structured and structured knowledge repositories

A first step towards ontologization can be achieved by integrating semi-structured content with structured resources, as provided, for instance, by existing taxonomies and semantic networks such as WordNet and Cyc.[14] To perform this integration the most accurate solution consists of using manual input from human experts (similarly to the general scenario of constructing knowledge resources, cf. Section 2). For instance, DBPedia [22] provides a knowledge base interlinked with various other Web resources – including, e.g., a manual mapping of Wikipedia entries to WordNet concepts – according to the Linked Data principle [21]. However, due to its high costs and limited scalability, a manual approach needs to be replaced or complemented with automatic methods, to which we now turn.

Flat representations: vector similarity and heuristic mapping. One of the first proposals for mapping Wikipedia to WordNet was presented by Ruiz-Casado et al. [173], who proposed associating Wikipedia pages with their most similar WordNet synsets on the basis of the similarity between their respective bags-of-words representations. Their method represents pages and synsets in a vector space model (using the pages' content and glosses, respectively) and computes their similarity using standard vector distance metrics such as, e.g., cosine or the dot product. An approach used in later work to construct BabelNet [138,139], a very large multilingual network combining WordNet and Wikipedia, achieves good performance by intersecting these bags-of-words. In contrast to using bags-of-words and vector space representations, Suchanek et al. [187] made use of the so-called 'most frequent sense' heuristic typically used in Word Sense Disambiguation (WSD, [132,133]). In order to build the YAGO ontology, they proposed unifying Wikipedia's category system with WordNet's taxonomy by linking each Wikipedia category to a WordNet synset. To this end, categories whose label consists of complex noun phrases (e.g.,

[14] Note that many of the approaches presented here are in fact part of larger proposals aimed at producing unified knowledge bases: consequently the discussion overlaps with many of the contributions presented later in Section 4.2.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

MATHEMATICIANS WHO COMMITTED SUICIDE) are first approximated by taking their lexical head (i.e., the most important noun found in the label, e.g., mathematician). Next, in order to establish the actual mapping, each category is associated with (the synset containing) the sense of its label: in the case that more than one such synset exists (i.e., the label consists of a polysemous word), a link is established with the sense which is most frequent in SemCor [119], a sense-labeled corpus. However, the most frequent heuristic can lead to many spurious mappings (e.g., linking the botanical Wikipedia category PLANTS to the 'industrial plant' sense of plant): consequently, later work from Ponzetto and Navigli showed how to improve this simple approach by means of a structure-based mapping algorithm [156].

Supervised and ensemble methods. As in the case of many complex tasks and applications, an alternative to unsupervised approaches consists of adopting supervised methods that make use of labeled data. These were explored for the mapping task by de Melo and Weikum in the construction of MENTA [114], a multilingual taxonomy which integrates WordNet and Wikipedia (Section 7). Their supervised model is induced from a set of manually-labeled mappings using features such as word-level information (term overlap between sets of synonyms and redirections, cosine similarity between the vector of glosses), as well as YAGO's most frequent sense heuristic (used here as a soft constraint, instead). Adopting a supervised approach for resource mapping has the advantage of yielding competitive results. However, as a trade-off, this approach is not applicable to arbitrary resources where no manually-labeled mappings are available for training.

Toral et al. [193] investigated a variety of methods for mapping Wikipedia categories to WordNet synsets, including a textual entailment system, knowledge-based semantic similarity (computed by applying Personalized PageRank [4] to the WordNet graph), as well as distributional methods [210]. While most methods achieve a performance comparable with a baseline mapping using SemCor's most frequent sense, substantial improvements are obtained by combining the output of these methods using a voting strategy or a supervised learner. Similar findings on using and combining Personalized PageRank with cosine similarity for the different, yet strongly related, task of mapping Wikipedia pages were later presented by Niemann and Gurevych [146].

Exploiting structure. Since Wikipedia provides semi-structured content, its structured features can also be exploited for the mapping task. Similarly to Ruiz-Casado et al., Medelyan and Legg [111] also integrated Cyc with Wikipedia by finding the semantically closest mappings. However, instead of relying on a flat vector representation, they quantified semantic distances using a relatedness measure computed on the Wikipedia hyperlink graph [120]. Toral et al. [195], instead, proposed linking Wikipedia to WordNet by finding categories and synsets which have the maximum overlap in terms of instances, that is, using structural evidence, but only from the lowest hierarchical level. A full-fledged graph-based method was presented by Ponzetto and Navigli [156], who associated categories from WikiTaxonomy [162] with those synsets which have the highest degree of structural overlap (computed against WordNet's taxonomy). Using a graph-based technique is found to improve over the most frequent sense heuristic by a large margin: the mapping, in turn, can be used to restructure the Wikipedia taxonomy, in order to improve its quality by increasing its degree of alignment with the reference resource (i.e., WordNet). The advantages of a structural approach to resource mapping were further investigated in the construction of BabelNet by Navigli and Ponzetto [139], who performed a side-to-side comparison of a bag-of-words method, originally developed in [138,157], with a graph-based approach. To provide a fair comparison, the two methods were evaluated within the same framework (i.e., to estimate the probability of a mapping), and the graph-based approach was consistently found to be superior in that it displayed a much higher recall for comparable precision.

Take-home message 2. Rather than representing knowledge sources that are mutually exclusive, *semi-structured and structured resources are complementary to each other*. In order to produce a unified, fully-ontologized resource, one needs to define a mapping procedure: the *best methods for this task* (i.e., highest-performing and requiring minimal supervision) are those which *take into account structural features* of the input resources, such as Wikipedia's hyperlink and category graphs, and WordNet's concept hierarchy.

4.2. Semantifying unstructured and semi-structured content: taxonomy and ontology induction

An alternative to integrating fully-structured knowledge sources with semi-structured information is to take semi-structured resources and build knowledge bases directly from them. In this section we look at approaches to taxonomy and ontology induction based on Wikipedia, and focus on the advantages of using semi-structured resources for this task.

Taxonomy induction: WikiTaxonomy. One of the first proposals for automatically acquiring a taxonomic resource from the category system of Wikipedia was presented by Ponzetto and Strube [159,155,161]. Their method starts with the category network 'as-is' (namely, with semantically unspecified relations between its categories, cf. Fig. 3(a)) and labels the relations between categories as *is-a* or *not-is-a*, thus producing a taxonomy as output (Fig. 3(b)). Semantic relations are established using a cascade of heuristics, which consider (i) the syntactic structure of the category labels, (ii) the topology of the category network, and (iii) the application of lexico-syntactic patterns for detecting subsumption and meronymy relations (based on the original work of Hearst [71] and Berland and Charniak [18], respectively). These three approaches are complementary: while specializations such as MATHEMATICIANS WHO COMMITTED SUICIDE *is-a* MATHEMATICIANS can be acquired

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

(a) Wikipedia category graph with unspecified semantic relations.

(b) Category graph retaining *isa* semantic relations only.

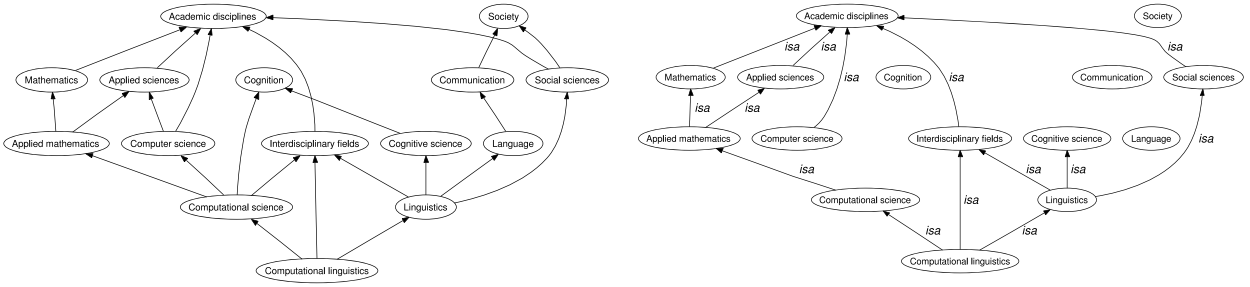

Fig. 3. Inducing taxonomic relations between categories in Wikipedia. Starting from the network of categories of Wikipedia, WikiTaxonomy is generated by traversing the network and deciding for each pair of categories whether the sub-category *isa* a super-category.

by means of a syntactic analysis of the category names, the last two types of method (i.e., (ii) and (iii)) generate more informative relations like, for instance, that MATHEMATICIANS WHO COMMITTED SUICIDE *is-a* SCIENTISTS WHO COMMITTED SUICIDE. The result is a large scale taxonomy containing hundreds of thousands of *is-a* relations. Evaluation of this automatically generated resource, named WikiTaxonomy, shows that it compares favorably in quality and coverage with existing manually created taxonomies such as WordNet and Cyc. In comparison with manually constructed resources, WikiTaxonomy contains a large amount of additional information such as, for instance, domain taxonomies covering very specialized areas in the fields of arts, business, fashion, etc. Furthermore, thanks to Wikipedia's common structure across different languages, the framework can be applied to wikipedias in languages other than English such as, e.g., German [84], to acquire taxonomic knowledge for other languages. Finally, while WikiTaxonomy is built upon a definition of the *is-a* relation which does not distinguish between subsumption (i.e., a concept-to-concept strict specialization/generalization) and instantiation (i.e., an entity-to-concept class membership) relations, this layer of information can be added later in a second stage – as shown by Zirn et al. [222], who also developed a cascade of heuristics to further specialize WikiTaxonomy's relations into *is-a* and *instance-of*.

Ontology acquisition: WikiNet, YAGO and DBPedia. Despite the fact that they typically form the backbone of structured repositories of knowledge, taxonomies contain only *is-a* relations. Instead, the range of semantic relations that can hold between concepts is far larger – e.g., meronymy (*part-of*) or even domain-specific relations (e.g., the *is-advisor-of* relation between a faculty academic and his or her students). To harvest such relations, Nastase and Strube [129] proposed building upon WikiTaxonomy and transforming it into a full-fledged semantic network by taking all category pairs labeled with a *not-is-a* relation, and specifying their semantics with more fine-grained relations such as *part-of*, *located-in*, etc. These are generated, again, by employing heuristics which look at the syntactic structure of the category labels, as well as at the connectivity in the category network. Part of this approach is also integrated in the construction of WikiNet described in this special issue in the paper by Nastase and Strube [130].

An alternative approach to building a full-fledged ontology from semi-structured content is used for YAGO [187]. YAGO consists of a wide-coverage semantic network which is automatically derived from Wikipedia and WordNet. Similarly to WikiTaxonomy and its extensions, it is built by means of a number of heuristics applied to semi-structured content from Wikipedia (e.g., infoboxes and categories), which is combined with taxonomic relations from WordNet. YAGO populates WordNet with millions of entities from Wikipedia by means of: (i) a simple, yet powerful, heuristic which assigns an *instance-of* relation to entities corresponding to Wikipedia pages and their (property-denoting) categories – e.g., Alan Turing *instance-of* BRITISH LONG-DISTANCE RUNNERS; (ii) a mapping from Wikipedia categories to WordNet synsets on the basis of SemCor's most frequent sense (Section 4.1). Other category-specific heuristics parse category labels and extract implicit relations between nominals – for instance, from the category 1954 Deaths one can derive that Alan Turing *died-in* 1954. Similar heuristics are also applied to the attribute-value pairs found within infoboxes in order to generate additional information (e.g., from `death_place=[[Wilmslow]]` it is possible to acquire Turing's place of death). All these relations are integrated within a unified knowledge base using an RDFS-like formalism, and are accessed by means of a dedicated query language. In recent years YAGO has enjoyed a wide interest and has been used for a variety of intelligent applications (e.g., IBM's Watson [57]): in this special issue, Hoffart et al. [73] present recent work which aims at taking this knowledge base to the next level by enriching it with spatial and temporal knowledge [72], in order that it can be applied to knowledge-intensive tasks requiring such knowledge types.

A parallel effort, similar in spirit to YAGO, is DBPedia [22], the goal of which is also to transform the semi-structured content of Wikipedia into a fully-structured representation using RDF, a Semantic Web language, as the underlying formalism. To this end, DBPedia builds upon a cascade of parsers to extract information from a variety of different elements found within Wikipedia pages, including redirects, inter-language links, categories, as well as infoboxes. The result is a very large ontology containing tens of millions of RDF statements. In contrast to YAGO, however, the focus of this project is on fully structuring information already present within pages, rather than acquiring novel, or abstracting existing, information.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

In addition, special focus is paid to linking the extracted knowledge with other resources, including YAGO itself and, among others, the CIA World Factbook,[15] Freebase[16] and OpenCyc.[17]

A heuristic renaissance. All approaches to taxonomy and ontology acquisition that we have seen in this section create structured knowledge repositories by means of heuristic approaches – arguably, an 'offbeat' tendency which contrasts with statistical approaches being the *de-facto* standard framework for work in knowledge acquisition. Each individual method is simple in nature, focuses on a specific characteristic of Wikipedia (e.g., the syntactic structure of the category labels, or the way categories and pages are connected within a hyperlink graph), and in combination these methods are able to produce very large resources with millions of concepts and semantic relations between them. In general, what is really interesting in this simple, yet powerful, heuristic approach is that it can be seen as a natural consequence of the availability of large amounts of high-quality semi-structured content. All the aforementioned heuristics, in fact, perform well because they start with content that is already meaningfully structured (i.e., Wikipedia's pages and categories), and need "only" to make it fully-structured by augmenting it with overt semantic information.

Take-home message 3. High-quality, *semi-structured content enables the acquisition of machine-readable knowledge on a large scale by means of heuristic methods* which essentially leverage regularities found within their shallow structure. That is, lightweight and scalable rule-based approaches can be devised to exploit the conventions governing the editorial base of collaboratively-generated resources, and capture large amounts of semantic information hidden within them.

5. Getting serious about semi-structured resources: exploiting collaboratively-generated knowledge

The availability of very large amounts of knowledge for many domains naturally invites, demands even, its use on a wide range of tasks. Accordingly, in this section, we present an overview of approaches aimed at leveraging knowledge from collaboratively-generated, semi-structured content for intelligent applications. To this end, we progressively look at phenomena of increasing complexity. We start in Section 5.1 with lexical semantic tasks concerned with modeling the meaning of words and phrases. Next, in Section 5.2 we look at high-end tasks which go beyond the word and sentence level, such as text categorization, Question Answering and Information Retrieval applications.

5.1. Understanding words and phrases

We first consider contributions which exploit semi-structured information for lexical semantic analysis. For this purpose, linguistic knowledge encoded within Wikipedia's structure has been found to be beneficial for several lexical understanding tasks, among which we focus here on Named Entity (NE) recognition and disambiguation, Word Sense Disambiguation (WSD), and computing semantic relatedness.

Named Entity Recognition. Named entity recognition (NER) aims at identifying mentions of named entities (NEs) in text. The task of recognizing NEs is a well-known problem in NLP (see [126] for a survey) and it has attracted a lot of interest in recent years, since information about NEs can benefit a variety of high-end applications such as, for instance, Web search [12]. Typically, the NER task consists of identifying mentions in running text and classifying them into coarse-grained semantic classes such as, for instance, PERSON, LOCATION and ORGANIZATION. Since Wikipedia is a very large repository of information that is focused primarily on NEs, it has been exploited for automatically generating labeled data and improving performance on NER.

A first proposal for harvesting structured information about entities was presented by Toral and Muñoz [194]. Their method extracts a gazetteer using the first sentence found on each Wikipedia page and relies on the fact that such sentences are typically definitional in nature.[18] For instance, given a sentence like:

(1) Alan Mathison Turing, OBE, FRS (23 June 1912–7 June 1954), was an English mathematician, logician, cryptanalyst, and computer scientist.

the method is able to extract the information that ALAN TURING is an instance of a PERSON. It achieves this by: (i) extracting the genus phrase from the sentence (mathematician, in this case), (ii) mapping it to its first sense in WordNet and (iii) following the *is-a* hierarchy until a synset denoting a named entity class is encountered (mathematician^1_n *is-a* scientist^1_n *is-a* person^1_n,[19] which in turn corresponds to the NE class PERSON). Similarly to the knowledge acquisition scenario from

[15] https://www.cia.gov/library/publications/the-world-factbook.
[16] http://www.freebase.com.
[17] http://www.opencyc.org.
[18] "The article should begin with a short declarative sentence, answering two questions for the nonspecialist reader: *What* (*or who*) *is the subject?* and *Why is this subject notable?*", extracted from http://en.wikipedia.org/wiki/Wikipedia:Writing_better_articles.
[19] We denote WordNet senses with w^i_p, namely the i-th sense of a word w with part of speech p.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

Section 4.2, this approach shows that large amounts of knowledge about NEs can be harvested by complementing semi-structured content with simple rules leveraging regularities in the annotations it contains. This semantic class information from Wikipedia's definitional sentences was later used by Kazama and Torisawa [85] as a feature for a supervised named entity recognizer. Improvements on a standard benchmarking dataset for this task [191] confirm the benefits of using gazetteer information mined from Wikipedia.

As we will see in this section in respect of a variety of tasks, the *leitmotif* of approaches to lexical semantic analysis exploiting Wikipedia is that they each leverage semi-structured data as semantic annotations. For instance, given a gazetteer where TURING is classified as an instance of PERSON, it is possible to label occurrences of internal links to the page ALAN TURING such as the following:

(2) The study of logic led directly to the invention of the programmable digital electronic computer, based on the work of mathematician *Alan Turing* and others.

as instances of an entity of type PERSON in context. Using Wikipedia as a source of NE annotated data has been investigated in a variety of papers, including those of Mika et al. [118], who use attribute-value pairs found in infoboxes, and Nothman et al. [147], who instead classify Wikipedia's articles into NE semantic classes and label occurrences of internal links with the class label (as in our example). The classification of a Wikipedia entry – i.e., identifying that the Wikipedia page ALAN TURING refers to an entity of class PERSON – can, in its turn, be performed in a number of ways. This includes using a manual list of keywords [171], a semi-supervised bootstrapping approach [147], or a supervised classifier using annotations from Wikipedia's bullet lists [206] or infoboxes [45]. The topic of automatically generating NE labeled data from Wikipedia and their extrinsic evaluation within a NER system is explored in detail in the article by Nothman et al. [148] included in this special issue.

Named Entity disambiguation. In order for the meaning of an NE to be fully modeled, it not only needs to be recognized in text, but also to be associated with its appropriate reference. For instance, the phrase Turing in these two sentences:

(3) A number of the university's faculty and students have also gained prominence in the field of computing sciences. Examples include [...] Ric Holt, co-creator of several programming languages, most notably *Turing*.
(4) *Turing* was played by Derek Jacobi, who also played in a 1996 television adaptation of the play.

refers to two different entities, namely the programming language vs. the British mathematician. The task of linking occurrences of ambiguous names with their corresponding NEs is referred to in the literature as named entity disambiguation.

Initial work on named entity disambiguation using Wikipedia as a source of entity-annotated data was presented by Bunescu and Paşca [30], who proposed viewing internal links as named entity annotations which define a link between a proper name in context (Turing, for instance) and the entity to which it refers, as given by its target Wikipedia page (e.g., TURING (PROGRAMMING LANGUAGE) or ALAN TURING). Bunescu and Paşca used this labeled data to build a model that maps each NE context and Wikipedia page to a vector space and links a proper name in context with the entity-denoting page which has maximal cosine similarity. Error analysis of the cosine-based ranking method shows that it suffers from the general limitations of the vector space model, i.e., it is not able to capture synonymy and does not cope with short or incomplete contexts. To provide a richer representation, feature vectors representing the contexts were therefore augmented with information quantifying the correlation between contexts and Wikipedia categories. This improved representation, combined with a disambiguation kernel, outperforms the vector space model and crucially demonstrates the beneficial effect of including category information from Wikipedia. The benefits of structural features mined from Wikipedia were further demonstrated in the subsequent work of Dredze et al. [51], where the performance of a supervised classifier is further boosted by including additional features derived from the Wikipedia hyperlink graph, such as the indegree/outdegree of an entity's Wikipedia page, as well as the page length in bytes.

Later contributions continued to build upon the idea of using Wikipedia as a repository of textual data containing a large quantity of annotated entities. Cucerzan [42] presented an unsupervised approach based on a vector space model, which jointly disambiguates all mentions of named entities within a document by assuming that entities referred to within a document must be related to each other. For instance, an occurrence of Java and Sun within the same document helps these two proper names to mutually disambiguate each other as referring, respectively, to the programming language and the company. This better disambiguation strategy is achieved, similarly to [30], by using information from Wikipedia categories to enrich the context vector representations of NE mentions. The benefits of a joint model that disambiguates all entities together were thereafter demonstrated with a wide range of algorithms, including simple hill-climbing and integer linear programming [94], as well as a graph-based algorithm combining syntactic and semantic compatibility information [74]. A detailed analysis of the task of entity disambiguation, including a systematic comparison of different state-of-the-art approaches and the most important factors affecting the performance of the systems, is presented in this special issue in the article by Hachey et al. [67].

Word Sense Disambiguation. While NE disambiguation is concerned with linking mentions of entities with their referents, Word Sense Disambiguation (WSD, [132,133]) aims at identifying the meaning of words in context. For this task knowledge

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

mined from semi-structured resources like Wikipedia has been shown to be beneficial both in a supervised and an unsupervised setting.

Mihalcea [117] proposed a method for automatically generating sense-tagged data which, similarly to work in NE recognition and disambiguation, views internal hyperlinks as linguistic annotations encoding lexico-semantic information, i.e., word sense annotations, in this case. For instance, two Wikipedia links anchored on the same word fellow are used in the following sentences to identify two different senses, the 'member of a learned society' vs. 'colleague'.

(5) In 1935, at the young age of 22, Turing was elected a *fellow* at King's on the strength of a dissertation in which he proved the central limit theorem.
(6) In 1941, Turing proposed marriage to Hut 8 co-worker Joan Clarke, a *fellow* mathematician and cryptanalyst, but their engagement was short-lived

By mapping Wikipedia articles to WordNet senses, hyperlinked text like these two example sentences can be transformed into standard sense-labeled data for use in WSD, which typically draws on WordNet as sense inventory [184,137]. Thus, Mihalcea was able to automatically create sense-annotated data on the basis of a manual mapping: these data were then provided as input to a machine learning WSD system for predicting the senses of words in unseen text. This approach was shown to improve the performance over the standard SemCor most frequent sense baseline [132], as well as a more complex approach based on the Lesk algorithm [99].

An alternative approach aimed at exploiting knowledge from Wikipedia for a knowledge-based WSD system was presented instead by Navigli and Ponzetto [139,157]. This approach maps Wikipedia pages to WordNet (cf. Section 4.1), enriches the latter with topical relations from the former, and uses this enriched knowledge base to perform graph-based WSD using context degree disambiguation [136]. Knowledge from Wikipedia is shown to be beneficial for knowledge-rich lexical disambiguation: when injected into WordNet, this knowledge, in fact, makes it possible for knowledge-based WSD algorithms to compete with the best performing approaches for this task, i.e., supervised systems, in a coarse-grained and multilingual setting, as well as to outperform them on domain-specific text.

Wikification: bringing entity and Word Sense Disambiguation together. An attempt to bring together the (different, yet complementary) entity and Word Sense Disambiguation tasks is provided by the so-called task of wikification. In a nutshell, wikification combines keyword extraction with lexical disambiguation: given an input document, a wikification system identifies the most important terms in the document and links (i.e., disambiguates) them to their appropriate entries within an external encyclopedic resource, i.e., typically Wikipedia. Csomai and Mihalcea [41] presented the first approach of this kind, in which an unsupervised keyword extraction algorithm using statistics drawn from Wikipedia's internal links is combined with the Lesk-like and supervised disambiguators presented by Mihalcea [117]. In later work by Coursey et al. [40], this method, named Wikify!, was applied in order to semantically tag arbitrary documents and automatically extract their most relevant topics using an unsupervised graph-based approach.

Wikification is a very active area of research, arguably because the ability to link documents can have a high impact on how vast amounts of online text, i.e. Web resources, can be structured automatically. The interest in this task is reflected by the popularity of community-driven shared tasks, such as the TAC Knowledge Base Population Track [110,80] and the NTCIR-9 Cross-Lingual Link Discovery Task,[20] which are two different kinds of wikification competitions: the former is aimed at benchmarking the performance of systems in linking mentions of entities in text to entries of a knowledge base built from Wikipedia, the latter, instead, is focused on the ability of systems to generate cross-lingual links between wikipedias in different languages. This heightened interest in wikification is also clearly apparent in two papers of this special issue. The Wikipedia Miner toolkit from Milne and Witten [122] makes the supervised wikification system originally presented in [121] freely available, while Tonelli et al. [192] present instead the Wiki Machine, a high-performance wikification system which is shown to outperform Wikipedia Miner thanks to a state-of-the-art kernel-based WSD algorithm [62].

Take-home message 4. *Wikipedia's internal links* can be viewed as *linguistic annotations of the meaning associated with nominal phrases* referring to word senses or named entities. These sense annotations, in turn, can be used to train state-of-the-art disambiguation models of lexical meaning in context, i.e., named entity, Word Sense Disambiguation or wikification systems.

Computing semantic relatedness. Recent years have seen a great deal of work on computing semantic relatedness, i.e., methods for automatically quantifying the strength of association between words. This is arguably because semantic relatedness provides a valuable model of semantic compatibility that can be embedded within a wide range of applications, including, among others, information retrieval [58], Word Sense Disambiguation [152], coreference resolution [158], short answer grading [124], and paraphrase detection [196].

Thanks to its multi-faceted semi-structure, Wikipedia has been drawn upon in a variety of different ways for computing semantic relatedness. Among the first such contributions was the WikiRelate! method developed by Ponzetto and

[20] http://ntcir.nii.ac.jp/CrossLink.

Strube [185,160], who used Wikipedia's category network to compute semantic relatedness by applying a variety of measures that had previously been developed for computing semantic relatedness exploiting lexical resources such as WordNet. The results show that, although Wikipedia outperforms WordNet on datasets modeling semantic relatedness such as the WordSimilarity-353 Test Collection [58], its performance on computing semantic similarity (i.e., a tighter notion of semantic compatibility) is nevertheless lower. This was because approaches for measuring semantic similarity that rely on lexical resources typically use paths based only on *is-a* relations [25]. Therefore, in later work [159,162] Ponzetto and Strube used their own WikiTaxonomy (Section 4.2) as a resource and applied the same measures as those used by WikiRelate! to compute semantic similarity using the *is-a* paths found in WikiTaxonomy. The results show that using the Wikipedia taxonomy makes it possible to outperform WordNet also on those datasets designed specifically for computing semantic similarity.

An alternative to using a structured representation such as Wikipedia's category network is offered by distributional methods [208,197]. Gabrilovich and Markovitch [59,60] introduced a method, called Explicit Semantic Analysis (ESA), which computes semantic relatedness on the basis of a vector space of concepts built from Wikipedia. The core idea behind ESA is that, given an input document, this document can be represented as a vector whose elements measure the strength of association between it and all other documents in a collection. Given a collection of documents such as Wikipedia, where each entry represents a disambiguated concept, the vector's elements can thus be taken to quantify the strength of association between the document's words and explicit concepts grounded within a semantic space. ESA offers a general retrieval model which has been analyzed in depth (see, e.g., [10]) and has been used for a variety of applications. The original results from Gabrilovich and Markovitch [59] showed that it enabled state-of-the-art performance on computing semantic relatedness between arbitrary text fragments (i.e., words or documents). Later work on ESA included its application to a variety of tasks such as text categorization [60] and semantic information retrieval [52], as well as extensions to include additional levels of information like, for instance, temporal information mined from Wikipedia's revision history [166].

One of the limitations of both WikiRelate! and ESA is that they do not exploit one of Wikipedia's richest features, namely, its hyperlink graph. Two approaches which fill this gap were presented in later work by Milne and Witten [120] and Yeh et al. [218]. Milne and Witten [120] compared a *tf*idf*-like measure computed on Wikipedia links with a more refined link co-occurrence measure modeled after the Normalized Google Distance [37]. Their findings indicate that exploiting the topology of the hyperlink graph yields a cheap, yet competitive, semantic relatedness measure. Yeh et al. [218], instead, exploited Wikipedia's structure by computing semantic relatedness using a random walk algorithm, i.e., Personalized PageRank [70], on a graph derived from Wikipedia's hyperlinks, infoboxes and categories. Their best results, obtained by combining information from categories and infoboxes, indicate that state-of-the-art performance can be achieved by leveraging multiple features from Wikipedia at the same time. However, using internal links only is found to degrade the overall performance: this is because these links represent many different types of semantic relation with different levels of strength of association – therefore, their richness needs to be combined with other resources and a robust weighting scheme. Using random walk algorithms for computing semantic relatedness is a very active area of research: when applied to Wikipedia, this approach is able, in fact, to exploit its structure (e.g., the internal link and category graphs) within a sound statistical framework. In this special issue, the paper by Yazdani and Popescu-Belis [216] explores the application of a random walk algorithm for computing semantic relatedness using Wikipedia.

Take-home message 5. *Wikipedia contains various different facets of semi-structured content*, including disambiguated concepts, internal links and categories, *all of which can be used for computing the strength of association between concepts and word senses* (i.e., semantic relatedness) on the basis of vector space and graph-based representations.

5.2. Language processing beyond the sentence level

Knowledge mined from collaboratively constructed resources can be used to perform tasks which go beyond the level of words and their surrounding local context (i.e., sentences, typically). In this section we review approaches which exploit knowledge from semi-structured resources for high-end tasks such as document clustering, text categorization, Question Answering and Information Retrieval applications.

Document clustering and text categorization. Documents can overlap in terms of domain, topic, entities they describe, their relations, etc. As a consequence, a first level of document understanding consists of either assigning documents a domain or topic label, or grouping them together on the basis of common features. These are two well-known tasks referred to in the literature as text categorization [182] and document clustering [33], respectively.

Knowledge mined from collaboratively-constructed resources has been successfully applied to both tasks. Banerjee et al. [13] show that performance on document clustering of news feeds can be improved by enriching their bag-of-words representation with the label of disambiguated concepts, i.e., the title of Wikipedia pages, which are related to the input short text. Hu et al. [78] proposed, instead, representing documents with a multi-level representation consisting of a content (i.e., bag-of-words) vector and concept vectors built from Wikipedia pages and categories. These vector representations are then combined and exploited to perform clustering by means of a variety of approaches (i.e., agglomerative vs. partitional) on the basis of their similarity. The results consistently show that information from Wikipedia categories gives the best improvements over the bag-of-words baseline, and that the best overall results can be achieved by combining it with page-level information.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

The results from both Banerjee et al. [13] and Hu et al. [78] show the usefulness of going beyond the simple bag-of-words model by including in the vector representation of texts features that explicitly model disambiguated concepts from Wikipedia. Indeed, this very same line of research has been successfully investigated for the task of text categorization, where improvements have been achieved by mapping documents onto semantic spaces in order to perform a more robust classification. One of the first such proposals was presented by Wang and Domeniconi [204], who represented documents by means of a semantic kernel which incorporates background knowledge from Wikipedia. Gabrilovich and Markovitch [60], instead, demonstrated the benefits of using ESA to represent documents in an explicit semantic space, prior to the classification step, whereas Wang et al. [205] explored the usefulness of expanding the bag-of-words representation with semantic relations from a thesaurus automatically constructed from Wikipedia. Recently, graph-based methods have been exploited by Navigli et al. [135] in order to perform document-level categorization jointly with term-level lexical disambiguation. In all cases, leveraging semantic features from Wikipedia is shown to boost performance in the classification task.

Take-home message 6. *Wikipedia*'s semi-structured content *can be used as a source of semantic information to enrich document representations and thus go beyond the simple bag-of-words model.* These semantically rich representations, in turn, are more effective on, and improve performance on, document-level tasks such as text clustering and classification.

Question Answering. The ability to answer questions is a natural benchmark for knowledge resources and methods which exploit them. Large amounts of curated knowledge, in fact, lie at the core of state-of-the-art Question Answering (QA) engines such as Wolfram Alpha[21] and TrueKnowledge,[22] while it is true to say that state-of-the-art QA approaches in general strive for wide-coverage content with a rich structure. Initial work on knowledge-based QA was concentrated on using manually-constructed knowledge repositories (e.g., WordNet-based query expansion [77]). However, the limited coverage of these resources soon led researchers to turn to the Web, whose huge amounts of textual data, although noisy, make it possible to acquire knowledge encoded within surface patterns on a large scale [168].

Wikipedia has been used in a variety of papers as a source of knowledge in order to overcome both the limited coverage of manual resources and the lower-than-human quality of knowledge automatically acquired from the Web. Although encyclopedic content is particularly beneficial for definitional questions (i.e., such as 'what is X?' or 'who is X?'), as highlighted by the findings of Lita et al. [100], Wikipedia has also been used for tackling other QA tasks. Ahn et al. [5] use it as a corpus in order to extract answers to both factoid questions and 'other' questions – i.e., interesting information about a topic from a target corpus that was not explicitly asked for in the factoid question [201]. Later work took this one step further by leveraging its structure with emphasis on the category system. Ahn et al. [6] used category labels as queries to find appropriate supporting documents in the target corpus for an input topic. Buscaldi and Rosso [31] matched input questions with Wikipedia's categories to collect candidate answers among their pages. An approach aimed, instead, at query expansion was presented by MacKinnon and Vechtomova [101], who used Wikipedia for Complex Interactive Question Answering [87], a QA task where systems are required to return information nuggets for an input topic, rather than entities or facts. Their approach leverages the anchor text of Wikipedia links as a source of expansion terms for improving the retrieval of relevant nuggets.

Recently, IBM's Watson [57] has renewed community-wide interest in QA by achieving human-level performance in the Jeopardy! game[23]: among its sources of content this state-of-the-art system combines the automatic corpus-based acquisition of text nuggets with wide-coverage information from Wikipedia contained in DBPedia and YAGO (Section 4.2). An in-depth analysis presented by Chu-Carroll and Fan [35] shows that the lexical answer types (namely the terms in the question that indicate the type of the entity to be returned) found in Jeopardy! questions are more fine-grained that those in TREC [202]: in order to fill the performance gap observed between datasets, they proposed mining candidate answers using anchor text and redirects metadata from Wikipedia documents. In this special issue, the paper on YAGO2 [73] includes an extrinsic evaluation of the quality of this knowledge base on the challenging task of answering spatio-temporal questions from the GeoCLEF 2008 GiKiP Pilot [177] and Jeopardy!.

Information Retrieval. Thanks to its repository of disambiguated concepts and its semi-structured content, Wikipedia naturally offers a semantically-rich environment for developing Information Retrieval (IR) applications beyond the bag-of-words model. The first contributions in this area have concentrated on using Wikipedia for the task of entity ranking. Entity ranking is a retrieval task where searchers are required to find documents representing entities of a certain type (e.g., computer scientists) that are relevant to an input query (e.g., researchers who worked on the halting problem). Evaluation of entity ranking systems has been conducted in the context of the INEX evaluation forum since 2006, using Wikipedia as the test collection [48,47]. As this is a task focused on entities and their semantic classes, most approaches have exploited category information from Wikipedia by either defining a similarity function between the categories associated with the queries and those of the retrieved entities, such as for instance category overlap [153], or by computing statistical associations of terms with category labels [83]. Category information can in turn be complemented with other structured content such as

[21] http://www.wolframalpha.com.
[22] http://trueknowledge.com.
[23] http://en.wikipedia.org/wiki/Jeopardy!.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

(a) Sentence compression

During the Second World War, Turing worked for the Government Code and Cypher School (GCCS) at Bletchley Park, Britain's codebreaking centre. For a time he was head of Hut 8, the section responsible for German naval cryptanalysis. He devised a number of techniques for breaking German ciphers, including the method of the bombe, an electromechanical machine that could find settings for the Enigma machine.
During the Second World War, Turing worked for the Government Code and Cypher School. For a time he was head of Hut 8, responsible for German naval cryptanalysis. He devised techniques for breaking German ciphers, including the method of the bombe.

(b) Lexical simplification

On 10 September 2009, following an Internet campaign, British Prime Minister Gordon Brown made an official public apology on behalf of the British government for the way in which Turing was treated after the war.
On 10 September 2009, following an Internet campaign, British Prime Minister Gordon Brown apologized on behalf of the British government for the way in which Turing was treated after the war.

(c) Discourse simplification

In July 1942, Turing devised a technique termed *Turingery* (or jokingly *Turingismus*) for use against the Lorenz cipher messages produced by the Germans' new *Geheimschreiber* (secret writer) machine.
In July 1942, Turing devised a technique termed *Turingery* (or jokingly *Turingismus*). Turingery was to be used against the Lorenz cipher messages. These messages were produced by the Germans' new *Geheimschreiber* (secret writer) machine.

(d) Paraphrasis

In 1928, Turing encountered Albert Einstein's work, and grasped it at a mere sixteen years of age, even extrapolating Einstein's Law of Motion from a text in which it was never made explicit.
In 1928, aged sixteen, Turing encountered Albert Einstein's work, and not only did he grasp it, but he extrapolated Einstein's Law of Motion from a text in which it was never made explicit.

Fig. 4. Diff between sentences from different revisions of the Wikipedia page for Alan Turing.

Wikipedia links, and used to perform text retrieval [153,46]. To overcome the fact that Wikipedia contains some entities that are very general (such as countries, for instance), retrieval can be coupled with an *idf*-like weighting scheme which penalizes over-linked entities [220]. Starting with entity linking, the paper by Kaptein and Kamps [82] in this special issue expands on the authors' previous work from [83] and shows how category information can also be exploited for other IR tasks such as ad-hoc retrieval.

A semantic IR system which leverages the conceptual space provided by Wikipedia has recently been presented by Egozi et al. [52]. Their approach consists of representing both documents and queries using concepts: this semantically-rich representation is achieved by projecting them into a conceptual vector space using Explicit Semantic Analysis [60]. The results indicate that, while using a semantic representation alone does not improve over a simple bag-of-words approach, its combination with a feature selection procedure is indeed able to achieve state-of-the-art results. In this special issue, the paper by Malo et al. [102] also tackles the problem of semantifying IR systems, however by concentrating, instead, on learning semantified queries. In this framework, named Wikipedia-based Evolutionary Semantics (Wiki-ES), wikified queries are learned using a variation of a genetic programming algorithm, and subsequently used to perform concept-based retrieval. Although ESA and Wiki-ES provide general retrieval frameworks, both approaches are based upon the semantic space of Wikipedia concepts. Thus we have yet further examples of the advantages of using a wide-coverage repository of disambiguated concepts.

Take-home message 7. The repository of disambiguated concepts found in *Wikipedia* (i.e., its articles) *provides a semantic space* into which documents and queries can be projected *in order to perform semantic retrieval* beyond the simple bag-of-words model.

6. Staying up-to-date: exploiting updated content from revision history

One of the most notable characteristics of Wikipedia is that, being an online collaborative medium, its content is continuously updated by human editors. Different versions of pages are stored and can be accessed and compared using a content management system. For instance, Fig. 4 shows the diff representation for fragments of different revisions of the Alan Turing page. Users can arbitrarily modify content by adding and/or removing text, as well as by rephrasing existing text. The sequence of different revisions of a Wikipedia article, called its revision history, provides a very large dataset of linguistic data, which can be used to model how documents evolve through time. This enables a range of NLP applications, and it is to these we now turn.

Rewriting tasks: sentence compression, text simplification and targeted paraphrasing. The ability to manipulate sentences in discourse includes monolingual text-to-text generation tasks such as sentence compression, text simplification and paraphrasing (among many others). Given an input sequence of words, sentence compression aims at finding a subset of this sequence which is both grammatical and meaning-preserving (Fig. 4(a)). Text simplification, on the other hand, can be viewed as a rewriting process whose output consists of more accessible – i.e., easier to understand – sentences

(Fig. 4(b)–(c)). Finally, paraphrasing consists of finding alternative ways of conveying the same information (Fig. 4(d)). While previous work on these tasks explored a variety of different approaches, including both statistical models using corpus data [90,109,39, *inter alia*] and hand-crafted rules [34,200], recent proposals have concentrated, instead, on leveraging Wikipedia's revision history as a source of data in order to automatically acquire sentence rewriting models.

Initial work in this line of research by Nelken and Yamangil [144,214] proposed creating a dataset of sentence pairs for sentence compression by iterating for each article over each temporally adjacent pair of revisions. For each such pair an edit-distance comparison is performed, treating each sentence as an atomic token. The dataset is then built by looking for all replacement operations of one sentence by another, and checking whether one sentence is a compression of the other. The resulting corpus is two orders of magnitude greater than the standard annotated dataset for this task (380,000 sentence pairs vs. the 1,067 from Knight and Marcu [90]): this much larger dataset enables a statistical system based on the noisy channel model [90] to achieve a higher compression and grammaticality rate.

In a similar vein, Yatskar et al. [215] and Woodsend and Lapata [211] created datasets of simplified word and sentence pairs, respectively, by finding reliable alignments between articles in the English and Simple English Wikipedia,[24] and mining the revision history of the latter. Both datasets are then used as training data to effectively learn statistical models of lexical and sentence simplification using a probabilistic model of edit operations [215], or by applying an integer linear programming model to select the most likely sentence simplification from the output of a quasi-synchronous grammar [211].

Finally, Max and Wisniewski [106] described a method for harvesting a corpus of sub-sentential paraphrases, which was later used by Bouamor et al. in [24] as a source of data to train a supervised model for validating candidate paraphrases using information from the Web. In line with findings for sentence compression and text simplification, Wikipedia's revision history was found to provide a viable source of linguistic data for training monolingual text-to-text generation models, in this case at the lower level of phrases.

Finding relevant terms within documents to improve search. Another application of revision history data is for modeling the evolution of documents through time, in order to identify important elements within them. For instance, Aji et al. [9] proposed including information from Wikipedia's revision history in order to score the relevance of terms within documents – i.e., under the assumption that important terms are introduced early in the document history and tend to be kept across revisions. Accordingly, their approach, named Revision History Analysis (RHA), scores terms on the basis of a frequency-based weighting scheme which captures the presence of a term across different document versions: this scoring function was later extended to model so-called 'burst', namely drastic content changes deriving from a sudden increase of popularity of a document. RHA is extrinsically evaluated as the term weighting component of two state-of-the-art IR models, namely BM25 and document language models [95], where it is shown to boost the performance on document search.

High-end semantic applications: Recognizing Textual Entailment. Document revisions also turn out to be a goldmine of data for the task of Recognizing Textual Entailment (RTE) [44,43]. RTE is a language understanding task which consists of recognizing whether the meaning of a text fragment (called the hypothesis) can be inferred from the meaning of another text fragment (called the premise). Although a wide range of different techniques for performing RTE have been developed over the years [11,17], a problem common to all of them has been the lack of large datasets and this, in turn, has limited the development of high-performance statistical models for the task. To overcome this issue, Zanzotto and Pennacchiotti proposed harvesting RTE corpora by mining Wikipedia's revision history [219]. Their approach starts by considering modified sentences in adjacent pairs of revisions as premise-hypothesis entailment pairs. A co-training approach [23] is then used to label these candidates as positive and negative instances, and thus iteratively provide additional data for an RTE supervised learner. The performance boosts on different benchmarking datasets indicate, yet again, the beneficial effects of using revision history data as training data.

Take-home message 8. Approaches using Wikipedia's revision history show that *versioned content can be exploited for the acquisition of statistical models* for different levels of linguistic structure. Applications which benefit from large amounts of diachronic information from Wikipedia range, in fact, from sub-sentential paraphrasing and sentence compression all the way up to tasks beyond the sentence level such as textual entailment, document search and discourse-level text simplification.

7. The tower of Babel: multilingual resources and applications

The Web is fundamentally a global and multilingual resource [63]. The existence of such enormous amounts of textual information in many languages virtually obliges researchers to develop novel methods for robust multilingual language processing. At the same time, vice versa, all this multilingual text represents a veritable goldmine of lexico-semantic knowledge in different languages which can be used for a wide range of applications. In this section, we look at how these two synergistic trends go hand in hand in the case of collaboratively-generated semi-structured resources.

[24] http://simple.wikipedia.org.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

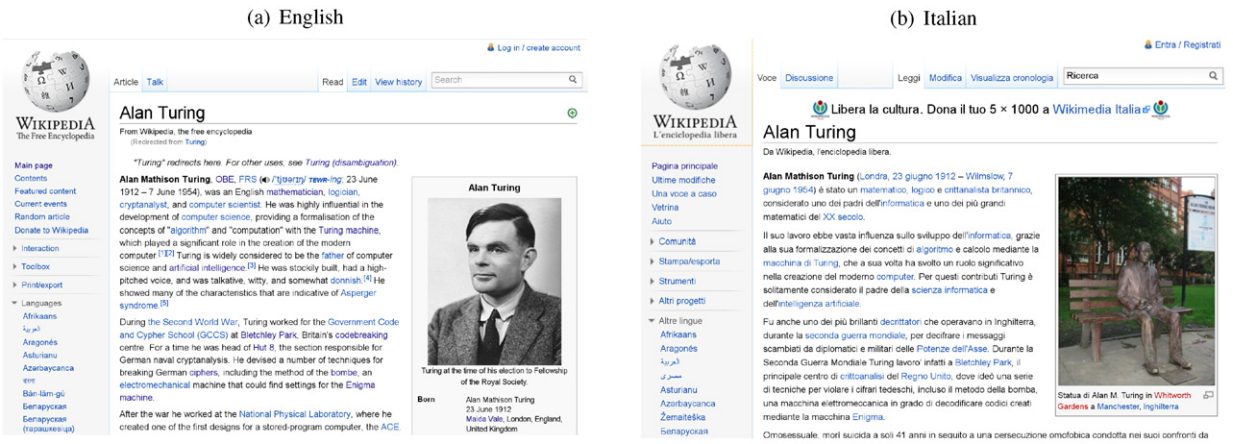

Fig. 5. English and Italian Wikipedia pages for ALAN TURING.

Multilinguality lies at the very core of collaborative resources such as Wikipedia. In fact, wikipedias in hundreds of different languages exist and these are highly interlinked with each other by means of so-called inter-language links. For instance, the English page for ALAN TURING (Fig. 5(a)) has an Italian counterpart (Fig. 5(b)): the two pages contain textual content and structural features such as infoboxes and categories which are comparable to each other (i.e., they overlap in meaning) despite being in different languages. Parallel multilingual structured resources could potentially benefit a variety of multilingual NLP applications: however, the limited availability of such content has, in the past, hindered research on knowledge-rich multilingual methods. In the following, we review approaches which overcome this problem by exploiting Wikipedia's content and structure across languages in order to perform multilingual knowledge acquisition (Section 7.1), and the application of this knowledge to different multilingual NLP tasks (Section 7.2).

7.1. Acquiring multilingual knowledge

Each of the different types of machine-readable repository of knowledge that we reviewed in Sections 3 and 4 – i.e., thesauri, semantic relations, taxonomies and ontologies – can be acquired from collaborative resources for languages other than English. Once again, the key idea is that information encoded within Wikipedia's shallow structure enables the acquisition of large amounts of multilingual lexico-semantic knowledge. In the approaches we will review below this is often achieved by leveraging Wikipedia's multilingual graph (i.e., inter-language links) and page content. So, what we are now going to look at is how repositories of knowledge of increasing depth and structure can be built from wikipedias in different languages.

Non-ontologized resources: parallel corpora and thesauri. An initial approach by Adafre and de Rijke [1] generated parallel corpora from Wikipedia by finding similar sentences in different languages which share links to the same concepts across languages. For instance, the two sentences:

(7) Alan Mathison Turing, OBE, FRS (23 June 1912–7 June 1954), was an English *mathematician*, *logician*, *cryptanalyst*, and *computer scientist*.
(8) Alan Mathison Turing (Londra, 23 giugno 1912–Wilmslow, 7 giugno 1954) è stato un *matematico*, *logico* e *crittanalista* britannico, considerato uno dei padri dell'informatica e uno dei più grandi matematici del XX secolo.

are deemed to be similar, since they contain the links in italics, which point to a description of the *same* concept in *different* languages, e.g., MATHEMATICIAN and MATEMATICO. Ye et al. [217], on the other hand, developed a graph-based approach for constructing a multilingual thesaurus. Their method represents a thesaurus as a weighted graph whose nodes include concepts and words in arbitrary languages and edges express relatedness between them (computed using a Dice coefficient on link co-occurrence statistics). This method is able to capture from the two sentences above, for example, that computer scientist and crittanalista are related words, since both occur with an instance of the same concept in different languages, namely MATHEMATICIAN and MATEMATICO. Basically, this approach builds on the method of Ito et al. [79], and extends it to include words in different languages from Wikipedia's inter-language links. The output is then extrinsically evaluated by performing multilingual query expansion on a variety of cross-lingual IR datasets from TREC [202].

Multilingual taxonomies and ontologies. In Section 4.1 we saw that high-quality knowledge acquisition can be attained by integrating wide-coverage content from Wikipedia with well-structured manual resources such as WordNet. An application of this line of research in a multilingual setting was presented by de Melo and Weikum in [113,114]. In [113] a Universal WordNet (UWN) is developed by automatically acquiring a semantic network for languages other than English. UWN is

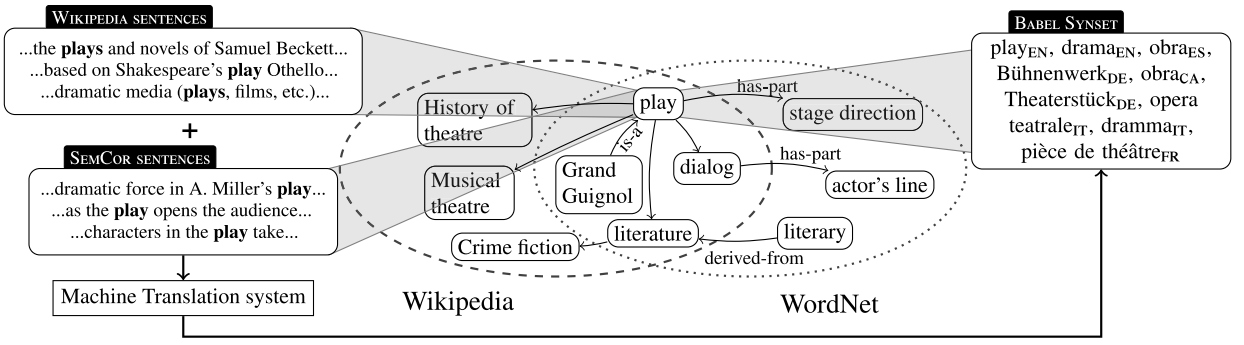

Fig. 6. An illustrative overview of BabelNet centered around the theatrical sense of play: relations between Babel synsets (i.e., multilingually lexicalized concepts) are obtained from hyperlinks between pages in Wikipedia (e.g., PLAY (THEATRE) links to MUSICAL THEATRE), as well as semantic relations between WordNet synsets (e.g., play^1_n *has-part* stage direction^1_n).

bootstrapped from WordNet and is built by collecting evidence extracted from existing wordnets, translation dictionaries, and parallel corpora. The result is a very large graph combining multilingual lexical information, i.e., multilingual word senses, with WordNet's taxonomic backbone. The same authors later presented an extension, named MENTA [114], which consists of a large-scale taxonomy of named entities and their classes also built from WordNet and Wikipedia. Both approaches focus crucially on building a multilingual taxonomy by (a) collecting large amounts of entities and translations from external resources (such as, e.g., online dictionaries and translations found in Wikipedia); (b) integrating these within the clean taxonomic structure of WordNet (UWN), which is further integrated with concepts from Wikipedia on the basis of a resource mapping phase (MENTA).

As was the case with taxonomy induction, Wikipedia's multilinguality also makes it possible to acquire multilingual ontologies. Efforts of this kind include monolingual resources in languages other than English, as well as lexical knowledge bases for a multitude of languages. An example of the former is the WOrdnet Libre du Français [175, WOLF], a French wordnet built using Wikipedia, among other resources, to provide a bilingual French–English semantic lexicon. An evaluation of WOLF shows that the straightforward use of inter-language links is able to provide accurate translations for multi-word expressions, a problem typically suffered by ontology learning approaches relying on parallel corpora. As was the case with monolingual approaches, substantial amounts of lexico-semantic information can be harvested by means of simple heuristics, which transform shallow annotations (inter-language links across wikipedias in different languages, in this case) into fully-structured knowledge – i.e., machine-readable disambiguated translations.

Nastase et al. [131] presented WikiNet, a very large multilingual semantic network built from Wikipedia. Similar in spirit to YAGO and DBPedia, WikiNet makes use of different facets of Wikipedia, including articles' internal and inter-language hyperlinks, infoboxes and categories. The concept inventory of WikiNet is constructed using all Wikipedia's pages (mostly, entities) and categories (typically referring to classes). Relations between concepts, instead, are harvested on the basis of a link co-occurrence analysis of Wikipedia's markup text – i.e., a semantically unspecified relation is assumed to exist between pairs of articles hyperlinked within the same window – and also collected from relations between categories generated using the heuristics developed by Ponzetto and Strube [162] and Nastase and Strube [129]. In this special issue, Nastase and Strube present in [130] an extension of their original work [131], and apply WikiNet to different NLP tasks, including computing semantic similarity and metonymy resolution.

All multilingual knowledge bases reviewed so far are primarily concerned with conceptual knowledge. At the lexical level, in fact, these approaches draw on Wikipedia's inter-language links (WikiNet), and additionally perform statistical inference, in order to produce a more coherent output (MENTA, see also [115]). A complementary contribution which is, instead, focused more on providing wide-coverage lexical knowledge for all languages, is BabelNet [138,139]. Similarly to other resources such as YAGO and UWN/MENTA, BabelNet aims at bringing together 'the best of both worlds' by combining very large numbers of entities found in Wikipedia with lexicographic knowledge from WordNet, by means of a high-performing unsupervised mapping framework (cf. Section 4.1). In addition, however, BabelNet also focuses specifically on providing high coverage for all languages thanks to Wikipedia's interlanguage links, and filling the so-called 'translation gaps' (i.e., missing translations from Wikipedia) by using statistical machine translation. Wikipedia's inter-language links are complemented with translations obtained from a state-of-the-art statistical machine translation system applied to sense-labeled data from SemCor [119], and Wikipedia itself. The result is a very large lexical multilingual knowledge base, overviewed in Fig. 6, which has been successfully applied to a variety of monolingual and cross-lingual lexical disambiguation tasks [139,141], and high-performance semantic relatedness [140].

Take-home message 9. *Wikipedia's multilinguality* – namely, the availability of interlinked wikipedias in different languages – *enables the acquisition of very large, wide-coverage repositories of multilingual knowledge*. Thanks to a common structure shared across languages, multilingual approaches can be developed which, in a similar fashion to their monolingual counterparts, make use of lightweight methods for exploiting semi-structured annotated data.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

7.2. Multilingual applications

As with the English monolingual scenario (Section 5), the ready availability of large amounts of multilingual knowledge naturally offers great potential for the development of multilingual applications. More specifically, the fact that lexicalizations for Wikipedia's disambiguated concepts are available in many different languages (via its inter-language links) makes it possible to develop applications which leverage this multilingual lexical knowledge for a variety of tasks. Richman and Schone [171], for instance, presented a method for automatically creating NE-labeled data for a variety of languages. The method starts by classifying Wikipedia's articles into NE semantic classes – e.g., Alan Turing is an instance of PERSON. Page classification is achieved on the basis of class-specific keywords (e.g., Alan Turing has categories 20TH-CENTURY MATHEMATICIANS and PEOPLE ASSOCIATED WITH BLETCHLEY PARK, which both vote for PERSON). Next, NE-annotated data are created by projecting the article's class onto occurrences of internal links to it (that is, all occurrences of internal links pointing to Alan Turing are labeled with PERSON, cf. Example 2, p. 12). Finally, by pivoting across languages on the basis of inter-language links, the method is able to assign the same NE class to Wikipedia articles in different languages and thus label non-English data with NE labels. For instance, given that Alan Turing has been classified using the English Wikipedia data as an instance of type PERSON, and that that page links to the homonym Italian Wikipedia page, it is possible to label the following internal link in the Italian Wikipedia with that same NE class:

(9) La Macchina di Turing come modello di calcolo è stata introdotta nel 1936 da *Alan Turing* per dare risposta all'Entscheidungsproblem (problema di decisione) proposto da Hilbert nel suo programma di fondazione formalista della matematica.

Richman and Schone used their method for the automatic creation of NE-annotated data for six languages other than English (French, Ukrainian, Spanish, Polish, Russian, and Portuguese) and then employed these as training data to bootstrap supervised NER systems in these languages. In this special issue, the paper by Nothman et al. [148] builds on Richman and Schone's work and their own previous work [147], in order to provide an in-depth analysis of methods for automatically creating multilingual NE corpora from Wikipedia and enabling NER in many different languages.

In general, as demonstrated by the multilingual propagation of NE labels, inter-language links essentially relate the *same* concept across *different* languages: this, in turn, makes it possible to straightforwardly project models built using the English Wikipedia data to other languages (and vice versa). Departing from this idea as a basis, Potthast et al. [165], Cimiano et al. [38] and Hassan and Mihalcea [69] each developed a cross-lingual version of ESA by mapping the concept vector representations from one Wikipedia article space to a Wikipedia in another language. As a result they were able to represent texts in arbitrary languages as vectors whose elements quantify the strength of association between words (in different languages) with (language-independent) concepts: this model is thus able to account, for instance, for the fact that the English computation and the Italian computazione are both strongly related with the named entity Alan Turing. Since vectors are concept-based, language-independent representations, similarity across languages can easily be computed by means of standard measures, e.g., cosine similarity. This, in turn, enables multilingual semantic applications such as cross-lingual search [165,38] and semantic relatedness [69]. An alternative to the distributional approach to semantic relatedness of Hassan and Mihalcea [69] was presented by Navigli and Ponzetto [140], who proposed computing semantic relatedness using the multilingual semantic information found in BabelNet. Their results show that graph-based methods are able to outperform distributional ones when provided with high-quality knowledge encoded within a multilingual semantic network, and that substantial improvements can be achieved by exploiting multilingual information from different languages jointly.

Take-home message 10. *Inter-language linking* relates descriptions of the same concept across wikipedias in different languages. This, in turn, *enables the projection of models across languages*, since the conceptual level provides a language-independent representation (i.e., a *conceptual interlingua*). As a result, information in arbitrary languages can be uniformly compared at the same semantic level for a variety of cross-lingual applications.

8. Conclusions

In this paper we presented a review of work from the recent literature on exploiting semi-structured, collaboratively built knowledge resources for AI and NLP, focusing primarily on Wikipedia as the main knowledge repository of this kind. The primary aims of this survey article were twofold: first, we wanted to provide readers with an up-to-date overview of the many different contributions that have been produced in this highly active area of investigation, so as to provide them with a comprehensive introduction to past and ongoing research directions; second, we were crucially interested in 'setting the stage' for the papers contained in this special issue.

The feedback we received from the research community for this special issue has been enormously positive, with 71 submissions received overall. These 71 papers all underwent an extremely rigorous selection process, carried out by an army of more than 150 reviewers who, together with the *Artificial Intelligence Journal* editors, made it possible for us to guarantee the highest standards. Thanks to this, the 10 papers that we were able to select for inclusion present top-level research on a wide spectrum of topics, ranging from computational neuroscience to the acquisition of structured knowledge on a large

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

scale, and its application to many different NLP and IR tasks. Accordingly, we organized the papers contained in this special issue on a thematic basis. The first two papers are devoted to acquiring structured knowledge from semi-structured sources [73,130]. These are followed by papers focused on Information Retrieval [102,82] and Natural Language Processing [67,148,216,192] applications. Finally, the last two papers present tools for working with semi-structured resources like Wikipedia [122], and its use to acquire computational semantic models of mental representations of concepts [154].

Back in 2006, some of us argued that 'we can only expect a bright future for automatic knowledge mining techniques with Wikipedia' [185]. Six years later it is clear that neither work on specific topics nor general interest in this research trend have diminished one bit. In this light, it would seem today that we can only expect an even brighter future, in terms not only of new methods, but also of novel tasks and innovative applications. In this sense the 'story so far' that we have presented here may well turn out to have been just the first step on a long line of path-breaking research exploiting semi-structured knowledge 'that anyone can edit'.

Acknowledgements

The last two authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234. Special thanks go to Jim McManus and all reviewers who contributed valuable feedback and ensured the papers contained in this special issue were of the highest quality. Finally, we thank the *Artificial Intelligence Journal* Editors-in-Chief, Tony Cohn, Rina Dechter and Ray Perrault, for their continued support throughout the preparation of this special issue.

References

[1] S.F. Adafre, M. de Rijke, Finding similar sentences across multiple languages in Wikipedia, in: Proceedings of the EACL-06 Workshop on New Text — Wikis and Blogs and Other Dynamic Text Sources, Trento, Italy, 4 April 2006.
[2] E. Adar, M. Skinner, D.S. Weld, Information arbitrage across multi-lingual Wikipedia, in: Proceedings of the Second ACM International Conference on Web Search and Data Mining, Barcelona, Spain, 9–12 February 2009, pp. 94–103.
[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne, Finding high-quality content in social media, in: Proceedings of the First ACM International Conference on Web Search and Data Mining, Palo Alto, Cal, 11–12 February 2008, pp. 183–194.
[4] E. Agirre, A. Soroa, Personalizing PageRank for Word Sense Disambiguation, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 30 March–3 April 2009, pp. 33–41.
[5] D. Ahn, V. Jijkoun, G. Mishne, K. Müller, M. de Rijke, S. Schlobach, Using Wikipedia at the TREC QA track, in: Proceedings of the Thirteenth Text REtrieval Conference, Gaithersburg, Md., 16–19 November 2004.
[6] K. Ahn, J. Bos, J.R. Curran, D. Kor, M. Nissim, B. Webber, Question answering with QED at TREC-2005, in: Proceedings of the Fourteenth Text REtrieval Conference, Gaithersburg, Md., 15–18 November, 2005.
[7] L. von Ahn, Games with a purpose, IEEE Computer 6 (2006) 92–94.
[8] L. von Ahn, L. Dabbish, Designing games with a purpose, Communications of the ACM 51 (2008) 58–67.
[9] A. Aji, Y. Wang, E. Agichtein, E. Gabrilovich, Using the past to score the present: extending term weighting models through revision history analysis, in: Proceedings of the Nineteenth ACM Conference on Information and Knowledge Management, Toronto, Ontario, Canada, 26–30 October 2010, pp. 629–638.
[10] M. Anderka, B. Stein, The ESA retrieval model revisited, in: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, Mass., 19–23 July 2009, pp. 670–671.
[11] I. Androutsopoulos, P. Malakasiotis, A survey of paraphrasing and textual entailment methods, Journal of Artificial Intelligence Research 38 (2010) 135–187.
[12] J. Artiles, S. Sekine, J. Gonzalo, Web people search: results of the first evaluation and the plan for the second, in: Proceedings of the 15th World Wide Web Conference, Beijing, China, 21–25 April 2008, pp. 1071–1072.
[13] S. Banerjee, K. Ramanathan, A. Gupta, Clustering short texts using Wikipedia, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23–27 July 2007, pp. 787–788.
[14] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, Open information extraction from the Web, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, 6–12 January 2007, pp. 2670–2676.
[15] M. Baroni, A. Kilgarriff, Large linguistically-processed web corpora for multiple languages, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3–7 April 2006, pp. 87–90.
[16] A. Belz, E. Kow, J. Viethen, A. Gatt, The GREC challenge 2008: overview and evaluation results, in: Proceedings of the Fifth International Natural Language Generation Conference, Salt Fork, Ohio, 12–14 June 2008.
[17] J. Berant, I. Dagan, J. Goldberger, Learning entailment relations by global graph structure optimization, Computational Linguistics 38 (2012) 73–111.
[18] M. Berland, E. Charniak, Finding parts in very large corpora, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Md., 20–26 June 1999, pp. 57–64.
[19] J.R.L. Bernard (Ed.), Macquarie Thesaurus, Macquarie, Sydney, Australia, 1986.
[20] J. Bian, Y. Liu, E. Agichtein, H. Zha, Finding the right facts in the crowd: factoid question answering over social media, in: Proceedings of the 17th World Wide Web Conference, Beijing, China, 21–25 April 2008, pp. 467–476.
[21] C. Bizer, T. Heath, T. Berners-Lee, Linked data – the story so far, International Journal on Semantic Web and Information Systems 5 (2009) 1–22.
[22] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, DBpedia – a crystallization point for the web of data, Journal of Web Semantics 7 (2009) 154–165.
[23] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the 11th Annual Conference on Learning Theory, Madison, Wisc., 24–26 July 1998, pp. 92–100.
[24] H. Bouamor, A. Max, G. Illouz, A. Vilnat, Web-based validation for contextual targeted paraphrasing, in: Proceedings of the ACL-11 Workshop on Monolingual Text-To-Text Generation, Portland, Oreg., 24 June 2011, pp. 10–19.
[25] A. Budanitsky, G. Hirst, Evaluating WordNet-based measures of semantic distance, Computational Linguistics 32 (2006) 13–47.
[26] P. Buitelaar, P. Cimiano, B. Magnini, Ontology learning from text: an overview, in: P. Buitelaar, P. Cimiano, B. Magnini (Eds.), Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, The Netherlands, 2005, pp. 1–10.

[27] P. Buitelaar, P. Cimiano, B. Magnini (Eds.), Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, The Netherlands, 2005.
[28] R. Bunescu, E. Gabrilovich, R. Mihalcea (Eds.), Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08, Chicago, Ill., 13 July 2008.
[29] R. Bunescu, E. Gabrilovich, R. Mihalcea, V. Nastase (Eds.), Proceedings of the Workshop on User-Contributed Knowledge and Artificial Intelligence: An Evolving Synergy at IJCAI-09, Pasadena, Cal., 13 July 2009.
[30] R. Bunescu, M. Paşca, Using encyclopedic knowledge for named entity disambiguation, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3–7 April 2006, pp. 9–16.
[31] D. Buscaldi, P. Rosso, Mining knowledge from Wikipedia from the question answering task, in: Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, 22–28 May 2006.
[32] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr., T.M. Mitchell, Toward an architecture for never-ending language learning, in: Proceedings of the 24th Conference on Artificial Intelligence, Atlanta, Georgia, 11–15 July 2010, pp. 1306–1313.
[33] C. Carpineto, S. Osiński, G. Romano, D. Weiss, A survey of web clustering engines, ACM Computing Surveys 41 (2009) 17:1–17:38.
[34] J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, J. Tait, Simplifying text for language-impaired readers, in: Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, Bergen, Norway, 8–12 June 1999, pp. 269–270.
[35] J. Chu-Carroll, J. Fan, Leveraging Wikipedia characteristics for search and candidate generation in question answering, in: Proceedings of the 25rd Conference on the Advancement of Artificial Intelligence, San Francisco, Cal., 7–11 August 2011, pp. 872–877.
[36] J. Chu-Carroll, J. Prager, An experimental study of the impact of information extraction accuracy on semantic search performance, in: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, Lisbon, Portugal, 6–9 November 2007, pp. 505–514.
[37] R. Cilibrasi, P.M.B. Vitányi, The google similarity distance, IEEE Transactions on Knowledge and Data Engineering 19 (2007) 370–383.
[38] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, S. Staab, Explicit vs. latent concept models for cross-language information retrieval, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, Cal., 14–17 July 2009, pp. 1513–1518.
[39] J. Clarke, M. Lapata, Global inference for sentence compression: an integer linear programming approach, Journal of Artificial Intelligence Research 31 (2008) 399–429.
[40] K. Coursey, R. Mihalcea, W. Moen, Using encyclopedic knowledge for automatic topic identification, in: Proceedings of the 10th Conference on Computational Natural Language Learning, Boulder, Col., 4–5 June 2009, pp. 210–218.
[41] A. Csomai, R. Mihalcea, Linking documents to encyclopedic knowledge, IEEE Intelligent Systems 23 (2008) 34–41.
[42] S. Cucerzan, Large-scale named entity disambiguation based on Wikipedia data, in: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, Prague, Czech Republic, 28–30 June 2007, pp. 708–716.
[43] I. Dagan, B. Dolan, B. Magnini, D. Roth, Recognizing textual entailment: rational, evaluation and approaches, Natural Language Engineering 15 (2009) i–xvii.
[44] I. Dagan, O. Glickman, B. Magnini, The PASCAL recognising textual entailment challenge, in: J. Quiñonero Candela, I. Dagan, B. Magnini, F. d'Alché Buc (Eds.), Machine Learning Challenges, in: Lecture Notes in Computer Science, vol. 3944, Springer, Heidelberg, 2006, pp. 177–190.
[45] W. Dakka, S. Cucerzan, Augmenting Wikipedia with named entity tags, in: Proceedings of the 3rd International Joint Conference on Natural Language Processing, Hyderabad, India, 7–12 January 2008, pp. 545–542.
[46] G. Demartini, C.S. Firan, T. Iofciu, R. Krestel, W. Nejdl, Why finding entities in Wikipedia is difficult, sometimes, Information Retrieval 13 (2010) 534–567.
[47] G. Demartini, T. Iofciu, A.P. de Vries, Overview of the INEX 2009 entity ranking track, in: S. Geva, J. Kamps, A. Trotman (Eds.), INEX, in: Lecture Notes in Computer Science, vol. 6203, Springer, 2009, pp. 254–264.
[48] G. Demartini, A.P. de Vries, T. Iofciu, J. Zhu, Overview of the INEX 2008 entity ranking track, in: S. Geva, J. Kamps, A. Trotman (Eds.), INEX, in: Lecture Notes in Computer Science, vol. 5631, Springer, 2008, pp. 243–252.
[49] P. Domingos, Toward knowledge-rich data mining, Data Mining and Knowledge Discovery 15 (2007) 21–28.
[50] D. Downey, O. Etzioni, S. Soderland, A probabilistic model of redundancy in information extraction, in: Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, 30 July–5 August 2005, pp. 1034–1041.
[51] M. Dredze, P. McNamee, D. Rao, A. Gerber, T. Finin, Entity disambiguation for knowledge base population, in: Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010, pp. 277–285.
[52] O. Egozi, S. Markovitch, E. Gabrilovich, Concept-based information retrieval using Explicit Semantic Analysis, ACM Transactions on Information Systems 29 (2011) 8:1–8:34.
[53] O. Etzioni, Search needs a shake-up, Nature 476 (2011) 25–26.
[54] O. Etzioni, M. Cafarella, D. Downey, A.M. Popescu, T. Shaked, S. Soderland, D.S. Weld, A. Yates, Unsupervised named-entity extraction from the web: an experimental study, Artificial Intelligence 165 (2005) 91–134.
[55] S. Faralli, R. Navigli, A new minimally-supervised framework for domain Word Sense Disambiguation, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, South Korea, 12–14 July 2012, pp. 1411–1422.
[56] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, MIT Press, Cambridge, Mass., 1998.
[57] D.A. Ferrucci, E.W. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J.W. Murdock, E. Nyberg, J.M. Prager, N. Schlaefer, C.A. Welty, Building Watson: an overview of the DeepQA project, AI Magazine 31 (2010) 59–79.
[58] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin, Placing search in context: the concept revisited, ACM Transactions on Information Systems 20 (2002) 116–131.
[59] E. Gabrilovich, S. Markovitch, Computing semantic relatedness using Wikipedia-based explicit semantic analysis, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, 6–12 January 2007, pp. 1606–1611.
[60] E. Gabrilovich, S. Markovitch, Wikipedia-based semantic interpretation for Natural Language Processing, Journal of Artificial Intelligence Research 34 (2009) 443–498.
[61] J. Giles, Internet encyclopedias go head to head, Nature 438 (2005) 900–901.
[62] C. Giuliano, A.M. Gliozzo, C. Strapparava, Kernel methods for minimally supervised WSD, Computational Linguistics 35 (2009) 513–528.
[63] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, J. McCrae, Challenges for the multilingual web of data, Journal of Web Semantics 11 (2012) 63–71.
[64] N. Guarino, C. Welty, Evaluating ontologies with OntoClean, Communications of the ACM 45 (2002) 61–65.
[65] I. Gurevych, T. Zesch (Eds.), Proceedings of the 2nd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources at COLING-10, Beijing, China, 28 August 2010.
[66] I. Gurevych, T. Zesch (Eds.), Proceedings of the 1st Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources at ACL-IJCNLP-09, Singapore, 7 August 2009.
[67] B. Hachey, W. Radford, J. Nothman, M. Honnibal, J.R. Curran, Evaluating entity linking with Wikipedia, Artificial Intelligence (2012), this issue, http://dx.doi.org/10.1016/j.artint.2012.04.005.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

[68] A.Y. Halevy, P. Norvig, F. Pereira, The unreasonable effectiveness of data, IEEE Intelligent Systems 24 (2009) 8–12.
[69] S. Hassan, R. Mihalcea, Cross-lingual semantic relatedness using encyclopedic knowledge, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 July 2009, pp. 1192–1201.
[70] T.H. Haveliwala, Topic-sensitive PageRank, in: Proceedings of the 11th World Wide Web Conference, Honolulu, Hawaii, 7–11 May 2002, pp. 517–526.
[71] M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: Proceedings of the 15th International Conference on Computational Linguistics, Nantes, France, 23–28 August 1992, pp. 539–545.
[72] J. Hoffart, F.M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, G. Weikum, YAGO2: exploring and querying world knowledge in time, space, context, and many languages, in: Proceedings of the 20th World Wide Web Conference, Hyderabad, India, 28 March–25 April 2011, pp. 229–232.
[73] J. Hoffart, F.M. Suchanek, K. Berberich, G. Weikum, YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia, Artificial Intelligence (2012), this issue, http://dx.doi.org/10.1016/j.artint.2012.06.001.
[74] J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum, Robust disambiguation of named entities in text, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, 27–29 July 2011, pp. 782–792.
[75] R. Hoffmann, C. Zhang, D.S. Weld, Learning 5000 relational extractors, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 286–295.
[76] T. Holloway, M. Bozicevic, K. Börner, Analyzing and visualizing the semantic coverage of Wikipedia and its authors: research articles, Complexity 12 (2007) 30–40.
[77] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, C.Y. Lin, Question answering in Webclopedia, in: Proceedings of the Ninth Text REtrieval Conference, Gaithersburg, Md., 13–16 November 2000, pp. 655–664.
[78] X. Hu, X. Zhang, C. Lu, E.K. Park, X. Zhou, Exploiting Wikipedia as external knowledge for document clustering, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 July–1 August 2009, pp. 389–396.
[79] M. Ito, K. Nakayama, T. Hara, S. Nishio, Association thesaurus construction methods based on link co-occurrence analysis for Wikipedia, in: Proceedings of the Seventeenth ACM Conference on Information and Knowledge Management, Napa Valley, Cal., 26–30 October 2008, pp. 817–826.
[80] H. Ji, R. Grishman, H.T. Dang, K. Griffitt, J. Ellis, Overview of the TAC 2010 knowledge base population track, in: Proceedings of the Third Text Analysis Conference, Gaithersburg, Md., 15–16 November 2010.
[81] D. Jurafsky, J.H. Martin, Speech and Language Processing, 2nd edition, Prentice–Hall, Upper Saddle River, N.J., 2008.
[82] R. Kaptein, J. Kamps, Exploiting the category structure of Wikipedia for entity ranking, Artificial Intelligence (2012), this issue, http://dx.doi.org/10.1016/j.artint.2012.06.003.
[83] R. Kaptein, P. Serdyukov, A.P. de Vries, J. Kamps, Entity ranking using Wikipedia as a pivot, in: Proceedings of the Nineteenth ACM Conference on Information and Knowledge Management, Toronto, Ontario, Canada, 26–30 October 2010, pp. 69–78.
[84] L. Kassner, V. Nastase, M. Strube, Acquiring a taxonomy from the German Wikipedia, in: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 26 May–1 June 2008.
[85] J. Kazama, K. Torisawa, Exploiting Wikipedia as external knowledge for named entity recognition, in: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, Prague, Czech Republic, 28–30 June 2007, pp. 698–707.
[86] F. Keller, M. Lapata, Using the web to obtain frequencies for unseen bigrams, Computational Linguistics 29 (2003) 459–484.
[87] D. Kelly, J.J. Lin, Overview of the TREC 2006 ciQA task, SIGIR Forum 41 (2007) 107–116.
[88] A. Kilgarriff, Googleology is bad science, Computational Linguistics 33 (2007) 147–151.
[89] A. Kilgarriff, G. Grefenstette, Web as corpus, Computational Linguistics 29 (2003) 333–348.
[90] K. Knight, D. Marcu, Statistics-based summarization – step one: sentence compression, in: Proceedings of the 17th National Conference on Artificial Intelligence, Austin, Tex., 30 July–3 August 2000, pp. 703–710.
[91] Z. Kozareva, E. Hovy, A semi-supervised method to learn and construct taxonomies using the web, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, Mass., 9–11 October 2010, pp. 1110–1118.
[92] E. Krahmer, K. van Deemter, Computational generation of referring expressions: A survey, Computational Linguistics 38 (2012) 173–218.
[93] M. Krötzsch, D. Vrandečić, M. Völkel, H. Haller, R. Studer, Semantic Wikipedia, Journal of Web Semantics 5 (2007) 251–261.
[94] S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti, Collective annotation of Wikipedia entities in web text, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 July–1 August 2009, pp. 457–466.
[95] J. Lafferty, C. Zhai, Document language models, query models, and risk minimization for information retrieval, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, La., 9–13 September 2001, pp. 111–119.
[96] M. Lapata, F. Keller, Web-based models for natural language processing, ACM Transactions on Speech and Language Processing 2 (2005) 1–31.
[97] D.B. Lenat, Cyc: a large scale investment in knowledge infrastructure, Communications of the ACM 38 (1995) 33–38.
[98] D.B. Lenat, R.V. Guha, Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project, Addison–Wesley, Reading, Mass., 1990.
[99] M. Lesk, Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, in: Proceedings of the 5th Annual Conference on Systems Documentation, Toronto, Ontario, Canada, pp. 24–26.
[100] L.V. Lita, W.A. Hunt, E. Nyberg, Resource analysis for question answering, in: Companion Volume to the Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004, pp. 162–165.
[101] I. MacKinnon, O. Vechtomova, Improving complex interactive question answering with Wikipedia anchor text, in: Proceedings of the 30th European Conference on Advances in Information Retrieval, Glasgow, U.K., 30 March–3 April 2008, pp. 438–445.
[102] P. Malo, P. Siitari, A. Sinha, Automated query learning with Wikipedia and genetic programming, Artificial Intelligence (2012), this issue, http://dx.doi.org/10.1016/j.artint.2012.06.006.
[103] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, 2008.
[104] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, Mass., 1999.
[105] Mausam, S. Soderland, O. Etzioni, D.S. Weld, K. Reiter, M. Skinner, M. Sammer, J. Bilmes, Panlingual lexical translation via probabilistic inference, Artificial Intelligence 174 (2010) 619–637.
[106] A. Max, G. Wisniewski, Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history, in: Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, 19–21 May 2010.
[107] J. McCarthy, Programs with common sense, in: Proceedings of the Teddington Conference on the Mechanization of Thought Processes, Her Majesty's Stationary Office, London, U.K., 1959, pp. 75–91.
[108] J. McCarthy, Epistemological problems of artificial intelligence, in: Proceedings of the 5th International Joint Conference on Artificial Intelligence, Cambridge, Mass., 22–25 August 1977, pp. 1038–1044.
[109] R. McDonald, Discriminative sentence compression with soft syntactic evidence, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3–7 April 2006, pp. 297–304.
[110] P. McNamee, H.T. Dang, Overview of the TAC 2009 knowledge base population track, in: Proceedings of the Second Text Analysis Conference, Gaithersburg, Md., 16–17 November 2009.
[111] O. Medelyan, C. Legg, Integrating Cyc and Wikipedia: folksonomy meets rigorously defined common-sense, in: Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08, Chicago, Ill., 13 July 2008, pp. 13–18.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

[112] O. Medelyan, D.N. Milne, C. Legg, I.H. Witten, Mining meaning from Wikipedia, International Journal of Human Computer Studies 67 (2009) 716–754.
[113] G. de Melo, G. Weikum, Towards a universal wordnet by learning from combined evidence, in: Proceedings of the Eighteenth ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009, pp. 513–522.
[114] G. de Melo, G. Weikum, MENTA: inducing multilingual taxonomies from Wikipedia, in: Proceedings of the Nineteenth ACM Conference on Information and Knowledge Management, Toronto, Ontario, Canada, 26–30 October 2010, pp. 1099–1108.
[115] G. de Melo, G. Weikum, Untangling the cross-lingual link structure of Wikipedia, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 844–853.
[116] C.M. Meyer, I. Gurevych, What psycholinguists know about chemistry: aligning Wiktionary and WordNet for increased domain coverage, in: Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, 8–13 November 2011, pp. 883–892.
[117] R. Mihalcea, Using Wikipedia for automatic Word Sense Disambiguation, in: Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, N.Y., 22–27 April 2007, pp. 196–203.
[118] P. Mika, M. Ciaramita, H. Zaragoza, J. Atserias, Learning to tag and tagging to learn: a case study on Wikipedia, IEEE Intelligent Systems 23 (2008) 26–33.
[119] G.A. Miller, C. Leacock, R. Tengi, R. Bunker, A semantic concordance, in: Proceedings of the 3rd DARPA Workshop on Human Language Technology, Plainsboro, N.J., 1993, pp. 303–308.
[120] D. Milne, I.H. Witten, An effective, low-cost measure of semantic relatedness obtained from Wikipedia links, in: Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08, Chicago, Ill., 13 July 2008, pp. 25–30.
[121] D. Milne, I.H. Witten, Learning to link with Wikipedia, in: Proceedings of the Seventeenth ACM Conference on Information and Knowledge Management, Napa Valley, Cal., 26–30 October 2008, pp. 1046–1055.
[122] D. Milne, I.H. Witten, An open-source toolkit for mining Wikipedia, Artificial Intelligence (2012), this issue, http://dx.doi.org/10.1016/j.artint.2012.06.007.
[123] M.L. Minsky, Semantic Information Processing, MIT Press, Cambridge, Mass., 1969.
[124] M. Mohler, R. Mihalcea, Text-to-text semantic similarity for automatic short answer grading, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 30 March–3 April 2009, pp. 567–575.
[125] R. Morante, C. Sporleder, Modality and negation: an introduction to the special issue, Computational Linguistics 38 (2012) 223–260.
[126] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Linguisticae Investigationes 30 (2007) 3–26.
[127] K. Nakayama, T. Hara, S. Nishio, Wikipedia mining for an association web thesaurus construction, in: Proceedings of the 8th International Conference on Web Information Systems Engineering, Nancy, France, 3–6 December 2007, pp. 322–334.
[128] V. Nastase, R. Navigli, F. Wu (Eds.), Proceedings of the Workshop on Collaboratively-Built Knowledge Sources and Artificial Intelligence at AAAI-10, Atlanta, Georgia, 11 July 2011.
[129] V. Nastase, M. Strube, Decoding Wikipedia category names for knowledge acquisition, in: Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence, Chicago, Ill., 13–17 July 2008, pp. 1219–1224.
[130] V. Nastase, M. Strube, Transforming Wikipedia into a large scale multilingual concept network, Artificial Intelligence (2012), this issue, http://dx.doi.org/10.1016/j.artint.2012.06.008.
[131] V. Nastase, M. Strube, B. Börschinger, C. Zirn, A. Elghafari, WikiNet: a very large scale multi-lingual concept network, in: Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, 19–21 May 2010.
[132] R. Navigli, Word Sense Disambiguation: a survey, ACM Computing Surveys 41 (2009) 1–69.
[133] R. Navigli, A quick tour of Word Sense Disambiguation, induction and related approaches, in: M. Bieliková, G. Friedrich, G. Gottlob, S. Katzenbeisser, G. Turán (Eds.), SOFSEM 2012: Theory and Practice of Computer Science, in: Lecture Notes in Computer Science, vol. 7147, Springer, Heidelberg, 2012, pp. 115–129.
[134] R. Navigli, G. Crisafulli, Inducing word senses to improve web search result clustering, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, Mass., 9–11 October 2010, pp. 116–126.
[135] R. Navigli, S. Faralli, A. Soroa, O.L. de Lacalle, E. Agirre, Two birds with one stone: learning semantic models for text categorization and word sense disambiguation, in: Proceedings of the Twentieth ACM Conference on Information and Knowledge Management, Glasgow, Scotland, U.K., 24–28 October 2011, pp. 2317–2320.
[136] R. Navigli, M. Lapata, An experimental study on graph connectivity for unsupervised Word Sense Disambiguation, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 678–692.
[137] R. Navigli, K.C. Litkowski, O. Hargraves, Semeval-2007 task 07: coarse-grained English all-words task, in: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007, pp. 30–35.
[138] R. Navigli, S.P. Ponzetto, BabelNet: building a very large multilingual semantic network, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 216–225.
[139] R. Navigli, S.P. Ponzetto, BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial Intelligence 193 (2012) 217–250, http://dx.doi.org/10.1016/j.artint.2012.07.001.
[140] R. Navigli, S.P. Ponzetto, BabelRelate! A joint multilingual approach to computing semantic relatedness, in: Proceedings of the 26th Conference on Artificial Intelligence, Toronto, Ontario, Canada, 22–26 July 2012, pp. 108–114.
[141] R. Navigli, S.P. Ponzetto, Joining forces pays off: multilingual joint Word Sense Disambiguation, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, South Korea, 12–14 July 2012, pp. 1399–1410.
[142] R. Navigli, P. Velardi, Learning Word-Class Lattices for definition and hypernym extraction, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 1318–1327.
[143] R. Navigli, P. Velardi, S. Faralli, A graph-based algorithm for inducing lexical taxonomies from scratch, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011, pp. 1872–1877.
[144] R. Nelken, E. Yamangil, Mining Wikipedia's article revision history for training computational linguistics algorithms, in: Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08, Chicago, Ill., 13 July 2008, pp. 31–36.
[145] D.P. Nguyen, Y. Matsuo, M. Ishizuka, Relation extraction from Wikipedia using subtree mining, in: Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence, Vancouver, B.C., Canada, 22–26 July 2007, pp. 1414–1420.
[146] E. Niemann, I. Gurevych, The people's web meets linguistic knowledge: automatic sense alignment of Wikipedia and WordNet, in: Proceedings of the 9th International Conference on Computational Semantics, Oxford, U.K., pp. 205–214.
[147] J. Nothman, T. Murphy, J.R. Curran, Analysing Wikipedia and gold-standard corpora for NER training, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 30 March–3 April 2009, pp. 612–620.
[148] J. Nothman, N. Ringland, W. Radford, T. Murphy, J.R. Curran, Learning multilingual named entity recognition from Wikipedia, Artificial Intelligence (2012), this issue, http://dx.doi.org/10.1016/j.artint.2012.03.006.
[149] B. O'Connor, R. Balasubramanyan, B.R. Routledge, N.A. Smith, From tweets to polls: linking text sentiment to public opinion time series, in: Proceedings of the 4th International AAAI Conference on Weblogs and Social, Media, Washington, D.C., 23–26 May 2010, pp. 122–129.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

[150] B. O'Connor, M. Krieger, D. Ahn, TweetMotif: exploratory search and topic summarization for Twitter, in: Proceedings of the 4th International AAAI Conference on Weblogs and Social, Media, Washington, D.C., 23–26 May 2010, pp. 384–385.
[151] M. Paşca, D. Lin, J. Bigham, A. Lifchits, A. Jain, Organizing and searching the world wide web of facts – step one: the one-million fact extraction challenge, in: Proceedings of the 21st National Conference on Artificial Intelligence, Boston, Mass., 16–20 July 2006, pp. 1400–1405.
[152] S. Patwardhan, S. Banerjee, T. Pedersen, Using measures of semantic relatedness for Word Sense Disambiguation, in: Proceedings of Computational Linguistics and Intelligent Text Processing, 4th International Conference, Mexico City, Mexico, 16–22 February 2003, pp. 241–257.
[153] J. Pehcevski, A.M. Vercoustre, J.A. Thom, Exploiting locality of Wikipedia links in entity ranking, in: Proceedings of the 30th European Conference on Advances in Information Retrieval, Glasgow, U.K., 30 March–3 April 2008, pp. 258–269.
[154] F. Pereira, M. Botvinick, G. Detre, Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments, Artificial Intelligence (2012), this issue, http://dx.doi.org/10.1016/j.artint.2012.06.005.
[155] S.P. Ponzetto, Creating a knowledge base from a collaboratively generated encyclopedia, in: Proceedings of the Doctoral Consortium at the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, N.Y., 22 April 2007, pp. 9–12.
[156] S.P. Ponzetto, R. Navigli, Large-scale taxonomy mapping for restructuring and integrating Wikipedia, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, Cal., 14–17 July 2009, pp. 2083–2088.
[157] S.P. Ponzetto, R. Navigli, Knowledge-rich Word Sense Disambiguation rivaling supervised systems, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 1522–1531.
[158] S.P. Ponzetto, M. Strube, Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution, in: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, New York, N.Y., 4–9 June 2006, pp. 192–199.
[159] S.P. Ponzetto, M. Strube, Deriving a large scale taxonomy from Wikipedia, in: Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence, Vancouver, B.C., Canada, 22–26 July 2007, pp. 1440–1445.
[160] S.P. Ponzetto, M. Strube, Knowledge derived from Wikipedia for computing semantic relatedness, Journal of Artificial Intelligence Research 30 (2007) 181–212.
[161] S.P. Ponzetto, M. Strube, WikiTaxonomy: a large scale knowledge resource, in: Proceedings of the 18th European Conference on Artificial Intelligence, Patras, Greece, 21–25 July 2008, pp. 751–752.
[162] S.P. Ponzetto, M. Strube, Taxonomy induction based on a collaboratively built knowledge repository, Artificial Intelligence 175 (2011) 1737–1756.
[163] H. Poon, P. Domingos, Unsupervised ontology induction from text, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 296–305.
[164] A.M. Popescu, M. Pennacchiotti, D. Paranjpe, Extracting events and event descriptions from Twitter, in: Companion Volume to the Proceedings of the 20th World Wide Web Conference, Hyderabad, India, 28 March–25 April 2011, pp. 105–106.
[165] M. Potthast, B. Stein, M. Anderka, A Wikipedia-based multilingual retrieval model, in: Proceedings of the 30th European Conference on Advances in Information Retrieval, Glasgow, U.K., 30 March–3 April 2008, pp. 522–530.
[166] K. Radinsky, E. Agichtein, E. Gabrilovich, S. Markovitch, A word at a time: computing word relatedness using temporal semantic analysis, in: Proceedings of the 20th World Wide Web Conference, Hyderabad, India, 28 March–25 April 2011, pp. 337–346.
[167] T. Rattenbury, N. Good, M. Naaman, Towards automatic extraction of event and place semantics from Flickr tags, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23–27 July 2007, pp. 103–110.
[168] D. Ravichandran, E. Hovy, Learning surface text patterns for a question answering system, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Penn, 7–12 July 2002, pp. 41–47.
[169] P. Resnik, N. Smith, The Web as a parallel corpus, Computational Linguistics 29 (2003) 349–380.
[170] M. Richardson, P. Domingos, Building large knowledge bases by mass collaboration, in: Proceedings of the 2nd International Conference on Knowledge Capture, Sanibel Island, Fl., 23–25 October 2003, pp. 129–137.
[171] A.E. Richman, P. Schone, Mining wiki resources for multilingual named entity recognition, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, Ohio, 15–20 June 2008, pp. 1–9.
[172] P.M. Roget, Roget's Thesaurus of English Words and Phrases, Penguin, Harmondsworth, U.K., 2000 (New ed./completely revised, updated and abridged by E.M. Kirkpatrick).
[173] M. Ruiz-Casado, E. Alfonseca, P. Castells, Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets, in: Advances in Web Intelligence, in: Lecture Notes in Computer Science, vol. 3528, Springer Verlag, 2005, pp. 380–386.
[174] M. Ruiz-Casado, E. Alfonseca, P. Castells, Automatic extraction of semantic relationships for WordNet by means of pattern learning from Wikipedia, in: Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems, Alicante, Spain, 15–17 June 2005, pp. 67–79.
[175] B. Sagot, D. Fišer, Building a free French WordNet from multilingual resources, in: Proceedings of the Ontolex 2008 Workshop, Marrakech, Morocco, 31 May 2008.
[176] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing and Management 24 (1988) 513–523.
[177] D. Santos, N. Cardoso, P. Carvalho, I. Dornescu, S. Hartrumpf, J. Leveling, Y. Skalban, GiKiP at GeoCLEF 2008: joining GIR and QA forces for querying Wikipedia, in: C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. Jones, M. Kurimo, T. Mandl, A. Peñas, V. Petras (Eds.), Evaluating Systems for Multilingual and Multimodal Information Access, in: Lecture Notes in Computer Science, vol. 5706, Springer, Heidelberg, 2008, pp. 894–905.
[178] P. Schmitz, Inducing ontology from Flickr tags, in: Proceedings of the WWW-06 Workshop on Collaborative Tagging, Edinburgh, Scotland, U.K., 22 May 2006.
[179] L.K. Schubert, Turing's dream and the knowledge challenge, in: Proceedings of the 21st National Conference on Artificial Intelligence, Boston, Mass., 16–20 July 2006, pp. 1534–1538.
[180] H. Schütze, Automatic word sense discrimination, Computational Linguistics 24 (1998) 97–124.
[181] H. Schütze, J.O. Pedersen, A cooccurrence-based thesaurus and two applications to information retrieval, Information Processing and Management 33 (1997) 307–318.
[182] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys 34 (2002) 1–47.
[183] B. Sigurbjörnsson, R. van Zwol, Flickr tag recommendation based on collective knowledge, in: Proceedings of the 17th World Wide Web Conference, Beijing, China, 21–25 April 2008, pp. 327–336.
[184] B. Snyder, M. Palmer, The English all-words task, in: Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3), Barcelona, Spain, 25–26 July 2004, pp. 41–43.
[185] M. Strube, S.P. Ponzetto, WikiRelate! Computing semantic relatedness using Wikipedia, in: Proceedings of the 21st National Conference on Artificial Intelligence, Boston, Mass., 16–20 July 2006, pp. 1419–1424.
[186] R. Studer, V.R. Benjamins, D. Fensel, Knowledge engineering: principles and methods, Data and Knowledge Engineering 25 (1998) 161–197.
[187] F.M. Suchanek, G. Kasneci, G. Weikum, YAGO: a large ontology from Wikipedia and WordNet, Journal of Web Semantics 6 (2008) 203–217.
[188] M. Surdeanu, M. Ciaramita, H. Zaragoza, Learning to rank answers to non-factoid questions from web collections, Computational Linguistics 37 (2011) 351–383.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002

[189] G. Szarvas, V. Vincze, R. Farkas, G. Móra, I. Gurevych, Cross-genre and cross-domain detection of semantic uncertainty, Computational Linguistics 38 (2012) 335–367.
[190] S. Tellex, B. Katz, J. Lin, A. Fernandes, G. Marton, Quantitative evaluation of passage retrieval algorithms for question answering, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Ontario, Canada, 28 July–1 August 2003, pp. 41–47.
[191] E.F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: language-independent named entity recognition, in: Proceedings of the 7th Conference on Computational Natural Language Learning, Edmonton, Alberta, Canada, 31 May–1 June 2003, pp. 142–147.
[192] S. Tonelli, C. Giuliano, K. Tymoshenko, Wikipedia-based WSD for multilingual frame annotation, Artificial Intelligence (2012), this issue, http://dx.doi.org/10.1016/j.artint.2012.06.002.
[193] A. Toral, O. Ferrández, E. Agirre, R. Muñoz, A study on linking Wikipedia categories to WordNet synsets using text similarity, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria, 14–16 September 2009, pp. 449–454.
[194] A. Toral, R. Muñoz, A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia, in: Proceedings of the EACL-06 Workshop on New Text — Wikis and Blogs and Other Dynamic Text Sources, Trento, Italy, 4 April 2006.
[195] A. Toral, R. Muñoz, M. Monachini, Named Entity WordNet, in: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 26 May–1 June 2008.
[196] G. Tsatsaronis, I. Varlamis, M. Vazirgiannis, Text relatedness based on a word thesaurus, Journal of Artificial Intelligence Research 37 (2010) 1–39.
[197] P.D. Turney, P. Pantel, From frequency to meaning: vector space models of semantics, Journal of Artificial Intelligence Research 37 (2010) 141–188.
[198] P. Velardi, R. Navigli, P. D'Amadio, Mining the web to create specialized glossaries, IEEE Intelligent Systems 23 (2008) 18–25.
[199] S. Verberne, L. Boves, N. Oostdijk, P.A. Coppen, What is not in the bag of words for *why*-QA? Computational Linguistics 36 (2010) 229–245.
[200] D. Vickrey, D. Koller, Sentence simplification for semantic role labeling, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, Ohio, 15–20 June 2008, pp. 344–352.
[201] E.M. Voorhees, Overview of the TREC 2004 Question Answering track, in: Proceedings of the Thirteenth Text REtrieval Conference, Gaithersburg, Md., 16–19 November 2004.
[202] E.M. Voorhees, D.K. Harman, TREC: Experiment and Evaluation in Information Retrieval, MIT Press, Cambridge, Mass., 2005.
[203] C. Wagner, Breaking the knowledge acquisition bottleneck through conversational knowledge management, Innovative Technologies for Information Resources Management 19 (2006) 70–83.
[204] P. Wang, C. Domeniconi, Building semantic kernels for text classification using Wikipedia, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, 24–27 August 2008, pp. 713–721.
[205] P. Wang, J. Hu, H.J. Zeng, Z. Chen, Using Wikipedia knowledge to improve text classification, Knowledge and Information Systems 19 (2009) 265–281.
[206] Y. Watanabe, M. Asahara, Y. Matsumoto, A graph-based approach to named entity categorization in Wikipedia using conditional random fields, in: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, Prague, Czech Republic, 28–30 June 2007, pp. 649–657.
[207] N.L. Waters, Why you can't cite Wikipedia in my class, Communications of the ACM 50 (2007) 15–17.
[208] J. Weeds, Measures and applications of lexical distributional similarity, Ph.D. thesis, Department of Informatics, University of Sussex, Brighton, U.K., 2003.
[209] G. Weikum, G. Kasneci, M. Ramanath, F. Suchanek, Database and information-retrieval methods for knowledge discovery, Communications of the ACM 52 (2009) 56–64.
[210] D. Widdows, K. Ferraro, Semantic Vectors: a scalable open source package and online technology management application, in: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 26 May–1 June 2008.
[211] K. Woodsend, M. Lapata, Learning to simplify sentences with quasi-synchronous grammar and integer programming, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, 27–29 July 2011, pp. 409–420.
[212] F. Wu, D. Weld, Automatically semantifying Wikipedia, in: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, Lisbon, Portugal, 6–9 November 2007, pp. 41–50.
[213] F. Wu, D. Weld, Open information extraction using Wikipedia, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 118–127.
[214] E. Yamangil, R. Nelken, Mining Wikipedia revision histories for improving sentence compression, in: Companion Volume to the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, 15–20 June 2008, pp. 137–140.
[215] M. Yatskar, B. Pang, C. Danescu-Niculescu-Mizil, L. Lee, For the sake of simplicity: unsupervised extraction of lexical simplifications from Wikipedia, in: Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, Cal., 1–6 June 2010, pp. 365–368.
[216] M. Yazdani, A. Popescu-Belis, Computing text semantic relatedness using the contents and links of a hypertext encyclopedia, Artificial Intelligence (2012), this issue, http://dx.doi.org/10.1016/j.artint.2012.06.004.
[217] Z. Ye, X. Huang, H. Lin, A graph-based approach to mining multilingual word associations from Wikipedia, in: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, Mass., 19–23 July 2009, pp. 690–691.
[218] E. Yeh, D. Ramage, C.D. Manning, E. Agirre, A. Soroa, WikiWalk: random walks on Wikipedia for semantic relatedness, in: Proceedings of the ACL-IJCNLP Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-4), Singapore, 7 August 2009, pp. 41–49.
[219] F.M. Zanzotto, M. Pennacchiotti, Expanding textual entailment corpora from Wikipedia using co-training, in: Proceedings of the 2nd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Beijing, China, 28 August 2010, pp. 28–36.
[220] H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, G. Attardi, Ranking very many typed entities on Wikipedia, in: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, Lisbon, Portugal, 6–9 November 2007, pp. 1015–1018.
[221] T. Zesch, C. Müller, I. Gurevych, Using Wiktionary for computing semantic relatedness, in: Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence, Chicago, Ill., 13–17 July 2008, pp. 861–867.
[222] C. Zirn, V. Nastase, M. Strube, Distinguishing between instances and classes in the Wikipedia taxonomy, in: Proceedings of the 5th European Semantic Web Conference, Tenerife, Spain, 1–5 June 2008, pp. 376–387.

Please cite this article in press as: E. Hovy et al., Collaboratively built semi-structured content and Artificial Intelligence: The story so far, Artificial Intelligence (2012), http://dx.doi.org/10.1016/j.artint.2012.10.002