Intelligenza Artificiale 7 (2013) 91–100
DOI 10.3233/IA-130051
IOS Press

Recent advancements in human language technology in Italy

Bernardo Magnini[a,*], Marco Baroni[b], Marcello Federico[a] and Roberto Navigli[c]
[a]*Fondazione Bruno Kessler, Trento, Italy*
[b]*University of Trento, Trento, Italy*
[c]*Sapienza University of Rome, Rome, Italy*

Abstract. This paper presents significant advancements in Human Language Technology fostered by Italian researchers in the last years. We report recent results in the following research directions: Distributional Semantics (i.e. the COMPOSES project), Multilingual Word Sense Disambiguation (the MultiJEDI project), Textual Semantic Inferences (the EXCITEMENT project), and Computer Assisted Translation (the MATECAT project). Key aspects that are common to such research initiatives include the fact that they are funded by the European Commission, under different grants, after having passed a very selective international competition; that they are led by Italian researchers, this way showing both the prominent role and the maturity of the Italian community on the international scene; and that they are currently running, indicating that the research topics of the projects are in the mainstream of Computational Linguistics.

Keywords: Human language technology, distributional semantics, multilingual word sense disambiguation, textual semantic inferences, computer assisted translation

1. Introduction

The Italian community on Human Language Technology (HLT) has been particularly active in the last years. Specific initiatives have been focusing on the Italian Language. Among the others, the recent volume on *The Italian Language in the Digital Age* [9] compares resources and tools available for Italian with corresponding resources and tools for other European languages, showing that Italian is quite well represented. Moreover, it is worth to mention the EVALITA initiative[1], which has run for three editions in 2007, 2009 and 2011, each culminated in a corresponding workshop (see the EVALITA 2011 proceedings, [22]), involving almost all the research groups in Italy, several participations outside Italy, and several companies. EVALITA is the reference evaluation campaign of both Natural Language Processing and Speech Technologies for the Italian language. The objective of the shared tasks proposed at EVALITA is to promote the development of language technologies for the Italian language, providing a common framework where different systems and approaches can be evaluated and compared in a consistent manner. As relevant results, EVALITA has both allowed to measure significant progress in the technology in a number of core tasks (e.g., among the others, part of speech tagging, named entity recognition, parsing) and made available an impressive amount of datasets for the tasks evaluated.

As an additional demonstration of the vivacity of the Italian community on HLT, a Special Issue on *Natural Language Processing in the Web Era* [4] has been recently published, collecting five papers that represent

*Corresponding author: Bernardo Magnini, Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy. E-mail: magnini@fbk.eu.

[1]http://www.evalita.it

significant topics about Natural Language Processing in the Web. Among these topics it is worth to mention the role of NLP for information access on linked open data, the automatic recognition of entities and their semantic coreference in the perspective of the Semantic Web, the use of NLP for accessing and combining available web services, as well as the extraction of content related to events mentioned in large document archives.

In this paper we present a selection of recent advancements in Human Language Technology, which can be ascribed as relevant achievements of the Italian community. The paper moves from the perspective of privileging the international visibility of research projects. Particularly, we report on four projects with the following characteristics: (i) the projects are funded by the European Commission, under different grants, after having passed a selective international competition; (ii) they are led by Italian researchers, this way showing a prominent role of the Italian community; (iii) the projects are currently running, so that research topics are in the mainstream of international research, and we can show preliminary results. Given the above criteria, we present the current progress of four projects: COMPOSES, MultiJEDI, EXCITEMENT and MATECAT.

COMPOSES (see Section 2) addresses compositionality in distributional semantics, a recent hot topic in Computational Linguistics. MultiJEDI (Section 3) is about multilingual word sense disambiguation, and aims to overcome the drawbacks of current supervised disambiguation systems. EXCITEMENT (Section 4) explores textual inferences, particularly textual entailment, under a modular component-based approach, aiming to realize an extendable open source platform. Finally, MATECAT (Section 5) aims to improve the integration of machine translation and human translation within the Computer Aided Translation framework.

2. COMPOSES: Compositionality in distributional semantics

COMPOSES (**COMP**ositional **O**perations in **SE**mantic **S**pace)[2] is a 5-year project financed by an European Research Council Starting Independent Research Grant that started in November 2011. The research team is coordinated by Marco Baroni at the Center for Mind/Brain Sciences of the University of Trento.

[2]http://clic.cimec.unitn.it/composes/

COMPOSES intends to fill the gap between two apparently orthogonal approaches to the study of meaning in natural language. *Formal semantics*, since the seminal work of Montague [25], has emphasized the *compositional* nature of language, observing how we can easily produce and understand an infinite number of sentences by combining the meanings of the words they include (you can understand the sentence *some dogs are pink* although you never read it before). Formal semantics has developed sophisticated tools to handle the logical scaffolding that enables meaning composition, but it has not come up with methods to build meaning representations for single words (to understand the sentence above you also need to now what *dogs*, *pink*, etc., mean) on the large scale required to account for human semantic knowledge (by the end of high-school, an average Western person is estimated to know the meaning of as many as 60,000 words [1]).

Distributional semantics [15, 38] is a (mostly computational) approach to meaning based on the idea that the latter can be characterized in terms of the set of contexts in which words occur in large collections of text, or corpora [24]. This approach has complementary strengths and weaknesses with respect to formal semantics. It can easily harvest effective meaning representations for thousands of words, but has no straightforward way to combine these representations to derive the meaning of phrases and sentences.

As in distributional semantics, the COMPOSES framework represents most content words (such as nouns) by vectors recording their corpus contexts. Implementing ideas from formal semantics, functional elements (such as adjectives, or determiners, like *the* and *many*) are represented by functions mapping from expressions of one type onto composite expressions of the same or other types (e.g., *yellow* is a function taking in input the meaning of a noun, such as *submarine*, and returning the meaning of another nominal element, *yellow submarine*). These composition operations, formalized as linear functions, are induced from corpus data by statistical learning of mappings from observed context vectors of input arguments to corpus-extracted context vectors of composite structures (e.g., *yellow* is represented by a weight matrix whose values are estimated by optimizing the mapping between corpus-extracted input-output vector examples, such as $\langle$*shirt, yellow shirt*$\rangle$, $\langle$*book, yellow book*$\rangle$, etc.). The general estimation framework based on corpus-extracted context vectors has been presented in [13].

Given the novelty of the approach, the COMPOSES project is also developing new evaluation frameworks:

On the one hand, taking inspiration from cognitive science and psycholinguistics, it designs elicitation methods to measure the perceived similarity and plausibility of sentences (elicited on a large scale by crowdsourcing). On the other, specialized entailment tests will assess the semantic inference properties of the corpus-induced system.

At the theoretical level, current results of COMPOSES include having demonstrated that compositional distributional representations of phrases can capture the meaning of *grammatical words*, such as determiners, that have been traditionally thought to be amenable of a logical treatment only [2, 5]. Research conducted within COMPOSES has also shown that the same compositional approach can be extended to word-internal derivation (deriving the meaning of *redo* from *re-* and *do*) [20]. The linear function method has been successfully extended to multi-argument functions, such as transitive verbs, modeled as third-order tensors that are multiplied by the two vectors representing object and subject as inputs, to return a vector representing the sentence they construct [18].

Perhaps the most important area the project is currently focusing on pertains to scaling up the system from phrases to sentences. Syntactic/semantic tree kernel methods along the lines of [8] are being explored to come up with a general estimate of sentence similarity that takes the similarity across distributional representations of all the phrases that compose them into account. Another area of active research pertains to the possibility of using compositional distributional representations not only to measure phrase or sentence *similarity*, but also *plausibility*, that is, being able to automatically tell that, although neither *coastal mosquito* nor *residential tomato* are attested in a very large corpus, the first phrase is expressing a more plausible concept than the second.

In terms of concrete deliverables, the COMPOSES project has already developed and made publicly available a toolkit to construct distributional representations of words and compose them [12]. Various small-scale data sets to test compositional models can be found on the project page (see in particular [36]), and a much larger data set comprised of 10,000 sentence pairs with sentence similarity and entailment judgments is currently under construction.

While COMPOSES focuses on basic research, and will mainly explore the theoretical implications of the combined distributional-formal approach for linguistics and cognitive science, to the extent that the project is successful, it promises to provide very useful tools for applied research as well. For example, the possibility to quantify how similar two phrases or sentences with arbitrary structures are paves the way to better information retrieval (querying search engines for meaningful *sentences*, and not just keyword strings). To the extent that distributional representations of sentences can also capture whether they entail each other, we expect them to serve as helpful cues in textual entailment systems (see Section 4).

One specific application where the project has already achieved good empirical results pertains to syntactic parsing, where plausibility measures applied to composed distributional vectors can help in solving difficult ambiguities, such as the correct bracketing of noun phrases [21]. For example, to decide if *miracle home run* should be parsed as *miracle (home run)* or *(miracle home) run*, the COMPOSES-based system measures the relative semantic plausibility of these phrases and their constituents – deciding, in this case, for the first grouping.

A broad overview of the theoretical and methodological underpinnings of COMPOSES, as well as of the research plan is presented in [3].

3. MultiJEDI: Multilingual word sense disambiguation

MultiJEDI (**Multi**lingual **J**oint word sens**E** **DI**sambiguation)[3] is a 5-year Starting Independent Research Grant funded by the European Research Council that started in February 2011. The research team is headed by Roberto Navigli at the Linguistic Computing Laboratory of the Department of Computer Science, Sapienza University of Rome.

The project aims to investigate new, groundbreaking directions in the field of *Word Sense Disambiguation* (WSD), the task of computationally determining the meaning of words in context [28, 29]. The key intuition underlying the project is that we now have the capabilities to transform multilinguality from an obstacle to Natural Language Understanding into a powerful catalyst for the task.

The most successful approaches to WSD are based on supervised machine learning. Given a word w, the typical process consists of learning a *word expert*, that is, a classifier that, given the context in which w occurs, associates a sense label with it – e.g. *spring* in the water sense in the sentence "*Spring* water can be found at

[3] http://multijedi.org/

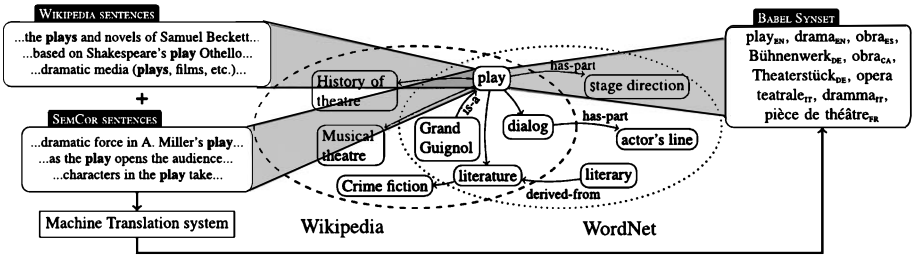

Fig. 1. An illustrative overview of BabelNet.

different altitudes." Typically, the sense inventory for a word is provided by a reference computational lexicon, that is, a highly structured lexical database (WordNet [16] being the most widespread). However, in order to attain state-of-the-art performance, supervised methods require large training sets tagged with word senses. It has been estimated that a high-accuracy supervised WSD system would probably need a text collection (i.e., a *corpus*) of about 3.2 million sense-annotated words [35]. At a throughput of one word per minute [14], this would require about 27 person-years of human annotation work. Even worse, all this effort would have to be repeated every time a new language became involved.

Therefore, in order to enable high-quality disambiguation in all languages, the MultiJEDI project focuses on advancing the knowledge-based paradigm in a multilingual setting. Knowledge-based WSD is based on the exploitation of wide-coverage lexical knowledge resources. It has been estimated that providing sufficiently large amounts of semantic relations has the potential to lead to considerable increases in the performance of knowledge-based systems, enabling them to compete with supervised approaches [31]. Thus, as a vital step towards its goal of performing disambiguation in any language, a major challenge for the project has been to create a large multilingual semantic network covering as many languages as possible. Our objective is to integrate encyclopedic and lexicographic knowledge from many languages within a unified resource, so as to create an "encyclopedic dictionary". To do this, we have developed an automatic graph-based algorithm which maps and integrates the largest dictionary of English, i.e. WordNet, and the most popular multilingual encyclopedia, i.e. Wikipedia. The result is a wide-coverage multilingual semantic network, named BabelNet [32], whose nodes are concepts (e.g., COMPUTER) and named entities (e.g., COMMODORE 64) and whose edges are lexical semantic relations (e.g., COMMODORE 64 *related-to* 8-BIT). Importantly, each node in the network contains a set of lexicalizations of the concept for different languages, e.g., $\{\text{play}_{\text{EN}}, \text{Theaterstück}_{\text{DE}}, \text{dramma}_{\text{IT}}, \text{obra}_{\text{ES}}, \ldots, \text{pièce de théâtre}_{\text{FR}}\}$. We call such multilingually lexicalized concepts *Babel synsets*. An overview of BabelNet is given in Fig. 1.

At the heart of BabelNet lies the use of semi-structured, collaboratively-created resources like Wikipedia, from which much-needed knowledge can be harvested [19] and linked to WordNet. Moreover, in order to increase lexical coverage in several languages, additional lexicalizations are obtained as a result of the application of a state-of-the-art machine translation system to sense annotated sentences (cf. Fig. 1).

The current version of the semantic network, i.e. BabelNet 2.0, covers 50 languages, including Italian, and is available online, together with API for its programmatic use.[4] Although even richer versions will become available within the next few years, the MultiJEDI project has already shown state-of-the-art results in monolingual and multilingual WSD using BabelNet for just four languages [34]. The proposed approach is graph-based and brings together the lexical knowledge from different languages by exploiting empirical evidence for the disambiguation from each of them, and then combining this information in a synergistic way: each language provides a piece of evidence for the meaning of a target word in context, and the integration of these various pieces enables them to constrain each other.

The availability of BabelNet made it possible to organize a task focused on multilingual WSD [30] in the context of the SemEval-2013 semantic evaluation competition.[5] The same dataset was annotated in five different languages, including Italian. Interestingly, several different knowledge-based systems participated in the task, whereas no supervised system did, probably due to the lack of training data for non-English

[4] http://babelnet.org

[5] http://www.cs.york.ac.uk/semeval-2013/task12/

languages. State-of-the-art systems achieved results ranging between 61% and 71% F1 depending on the language.

BabelNet has paved the way not only towards multilingual WSD, but also towards multilingual semantic relatedness [33] and the large-scale harvesting of semantic predicates [17]. Moreover, the work conducted on BabelNet and multilingual WSD within the MultiJEDI project helped establish relationships with renowned groups, such as that of Hans Uszkoreit and Feiyu Xu (Saarland University and DFKI), with whom research in knowledge-based Information Extraction is ongoing in the context of a Google Focused Research Award. Initial results are extremely encouraging: using a graph-based algorithm, BabelNet has been shown to considerably improve the precision of a non-semantic relation extraction system, while at the same time retaining high recall [26]. More applications are currently under study, including multilingual domain-focused and Open Information Extraction, with the aim of labeling and semantifying the relations which connect our multilingual synsets [27].

4. EXCITEMENT: Textual semantic inferences

Excitement (Exploring Customer Interactions through Textual Entailment)[6] is a 3-year research project (1/2012-12/2014) funded by the European Commission under FP7. The project consortium is led by a company, NICE (Israel), with FBK as a scientific coordinator, and includes the University of Bar Ilan (Israel), DFKI (Germany), University of Heidelberg (Germany), and two companies, Almawave (Italy) and OMQ (Germany).

The project addresses the issue of identifying semantic inferences between text units [10], which is a major underlying language processing task, needed in practically all text understanding applications. Particularly, we focus on textual entailment recognition, i.e. the task of deciding, given two text fragments, whether the meaning of one text is entailed (can be inferred) from another text [**?**]. This task captures generically a broad range of inferences that are relevant for multiple applications. A necessary step in transforming textual entailment from a theoretical idea into an active empirical research field was the introduction of benchmarks and an evaluation forum for entailment systems. Dagan, Glickman and Magnini initiated in 2004 a series of contests under the PASCAL Network of Excellence, known as *The PASCAL Recognising Textual Entailment Challenges* (*RTE* in short). These contests provided researchers concrete datasets on which they could evaluate their approaches, as well as a forum for presenting, discussing and comparing their results. The RTE datasets are freely available also for non-RTE participants, so as to further facilitate research on textual entailment.

While such inferences are broadly needed, there are currently no generic semantic engines or platforms for broad textual inference. The primary scientific motivation for the EXCITEMENT project is to change this ineffective state of affairs and to offer an encompassing open source platform for textual inference. On the industrial side, EXCITEMENT is focused on the text analytics market and follows the increasing demand for automatically analyzing customer interactions, which today cross multiple channels including speech, email, chat and social media.

There are two interleaved high-level goals for this project, which would yield two corresponding outcomes. The first is to set up, for the first time, a generic architecture and a comprehensive implementation for a multilingual textual inference platform and to make it available to the scientific and technological communities. To a large extent, the idea is to follow the successful experience of the Moses open source environment for machine translation, which has been making substantial impact on research in that field. This will enable developers of many text-processing applications to leverage the platform and boost their semantic inference capabilities. It will also provide developers of inference technology an effective environment for implementing and evaluating their components, and an easy entry-point for research in this field.

The second goal of the project is to develop a new generation of inference-based industrial text exploration applications for customer interactions, which will enable businesses to better analyze and make sense of their diverse and often unpredicted client content. These goals will be achieved for three languages, i.e. English, German and Italian, and for three customer interaction channels, i.e. speech (transcriptions), email and social media.

The expected impact of EXCITEMENT is driven by the prospect to lay new grounds for powerful textual inference technology. On the scientific side, the realization of the generic textual entailment paradigm within an encompassing open platform would end the current state of affairs in applied semantics technology, which, for a long time was lacking a feasible unifying driv-

[6] http://www.excitement-project.eu

ing framework. On the industrial side, providing new inference capabilities will open new horizons for text analytics in general, and customer interaction analytics in particular, which will enable businesses to better harness the value of their customer inputs and thus increase their competitiveness.

Altogether, the expected outcome of EXCITEMENT ranges from new scientific insights all the way to novel practical technology in the hands of European developers and end-users. The consortium dedication to the open source platform, shared resources and transparent scientific dissemination will ensure that the results strengthen the R&D base in applied semantics and text exploration: from university students to commercial developers, all types of both academic and industrial players will benefit from the proposed project.

A major result of the project so far (i.e. month 18) is the first release of the EXCITEMENT Open Platform (EOP-1.0.1). We started with three entailment systems developed by the project partners: BIUTEE from Bar Ilan, TIE from DFKI, and EDITS from FBK, which have been migrated in the EOP architecture. The platform aims to automatically check for the presence of entailment relations among texts. It is based on a modular architecture and provides support for the development of algorithms that are language independent to a high degree. Thus, it allows developers and users to combine linguistic pipelines, entailment algorithms and linguistic resources within and across languages with as little effort as possible. As an example, a classification-based entailment algorithm can use, both separately and in combination, the results of a distance component and the results of a BoW component, with the possibility to use lexical resources (e.g. wordnets) in different languages. The result is an ideal software environment for experimenting and testing innovative approaches for textual inferences. The EOP is distributed as open source software[7] and its use is open both to users who are interested in integrating inference technology in applications, and to developers who are willing to extend the current functionalities.

The platform includes the following main characteristics.

Multilinguality. Given an input Text-Hypothesis pair(s) in one of the three languages of the project (i.e. English, German, Italian), the EOP allows to produce corresponding entailment judgments (i.e. *Entailment*, *No-Entailment*).

[7] http://hltfbk.github.io/Excitement-Open-Platform/

Component-Based approach. The EOP conforms to the component-based approach of the project. Existing functionalities of the existing entailment engines are adapted to the blocks of the EOP architecture, which includes: pipeline, entailment decision algorithms (EDAs), knowledge components, distance components, resources.

Learning capacity. The EOP is able to build models on training data, which are then evaluated on corresponding test data. For this purpose we use the RTE-3 dataset in the three languages.

Focus on lexical entailment. We start from entailment issues that are based on lexical phenomena in the three languages, for which all the academic partners have experience. For each language we individuate lexical resources contributing to entailment (e.g. WordNet, Wikipedia), and build corresponding knowledge components, distance components, and entailment algorithms. Typical phenomena that are responsible for lexical entailment include: synonymy, hyperonymy, antonymy, morphological derivations, named entities variations, acronyms, abbreviations, world knowledge about people and locations, etc. Although the focus of the first version of the EOP is lexical entailment, we plan to rapidly move to syntactic phenomena thanks to the components and entailment algorithms developed in BIUTEE and TIE.

Relation with the project applicative scenarios. The first prototype already serves some of the requirements of the two industrial applications foreseen in the project: entailment-based graph exploration and entailment-based retrieval.

The current results of the project include different levels of interoperability among the components of the platform:

- *Linguistic annotation interoperability*. Linguistic annotations for different languages can be used by the same entailment algorithm. This interoperability is achieved through the use of the same format (i.e. UIMA-CAS) and the same set of semantic types by the linguistic pipelines for the three languages.
- *Resource interoperability*. Different linguistic resources can be used by the same entailment algorithm. This interoperability is achieved through the lexical component interface, which basically assumes that knowledge extracted from different

resources is represented as entailment rules. As an example, although the Italian WordNet and the Italian Wikipedia are stored in completely different formats, their relevant content is managed by the lexical component interface as entailment rules, allowing a single entailment algorithm (e.g. the edit distance entailment algorithm) to use both in a completely transparent way.
- *Entailment algorithm interoperability*. Different entailment algorithms can use the same distance component. This interoperability is achieved through a strict separation of the algorithm taking the entailment decision from the algorithm that calculates the distance between Text and Hypothesis in a pair (i.e. the distance component). As an example, both the Edit Distance entailment algorithm (from Edits) and the classification-based entailment algorithm (from TIE) can take advantage of the result of the Distance component (from Edits).
- *Component interoperability*. Different components can be used by the same entailment algorithm. As in the previous case, this interoperability is achieved through a strict separation of the algorithm taking the entailment decision from the algorithm that calculates the distance between T and H in a pair. As an example the classification-based entailment algorithm (from TIE) can use, both separately and in combination, both the results of the distance component (from Edits) and the results of the BoW component (from TIE).

Given the complexity of the project, in term of the number of components and their different license types, the EOP distribution has required particular attention as far as both distribution and installation procedures are concerned, and the management of software packages using different programming languages.

5. MATECAT: Computer assisted translation

MateCat (Machine Translation Enhanced Computer Assisted Translation) is a 3-year research project (11/2011-10/2014) funded by the European Commission under FP7. The project consortium is lead by FBK and includes The University of Edinburgh (United Kingdom), Université du Maine (Le Mans, France), and Translated Srl (Rome, Italy).

The objective of MateCat is to improve the integration of machine translation and human translation within the so-called computer aided translation (CAT) framework. CAT tools represent nowadays the dominant technology in the translation industry. They provide translators with text editors, that can manage several document formats and suitably arrange their content into segments – i.e. sentences–, ready to be translated. Most importantly, CAT tools provide access to dictionaries, to translation memories (TMs), and more recently to machine translation (MT) engines. A TM is basically a repository of translated segments. During translation, the CAT tool queries the TM to search for exact or fuzzy matches of the current source segment. These matches are proposed to the translator as translation suggestions. Once a segment is translated, its source and target texts are added to the TM for future queries. Recently, when no good matches are found in the TM, suggestions from an MT engine are also supplied to the translator.

Recent studies have shown that post-editing suggestions from a statistical MT engine can substantially improve productivity of professional translators. MateCat leverages the growing interest and expectations in statistical MT by advancing the state of the art along directions that will hopefully accelerate its adoption by the translation industry. In particular, MateCat investigates the integration of MT into the CAT working process along three main research directions:

- *Self-tuning MT*, i.e. methods to train statistical MT engines for specific domains or translation project;
- *User adaptive MT*, i.e. methods to quickly adapt statistical MT from user corrections and feedback.
- *Informative MT*, i.e. supply users with additional information to enhance their productivity and work experience.

These new MT functionalities will be integrated in a new Web-based CAT tool: the *MateCat Tool*. The MateCat Tool will provide both a professional work environment integrating advanced MT functionalities, and a research platform to run MT post-editing experiments and to measure user productivity.

Progress of the project will be measured through field tests evaluating the utility and usability of the new MT functionalities. Field tests will be carried out with professional translators performing real translation projects with the MateCat Tool. The considered translation directions are English-Italian, English-German, English-French, and English-Spanish. Experimental evaluations are performed on two domains: legal documents produce by the European Commission, and technical manuals produced by IT companies. After one

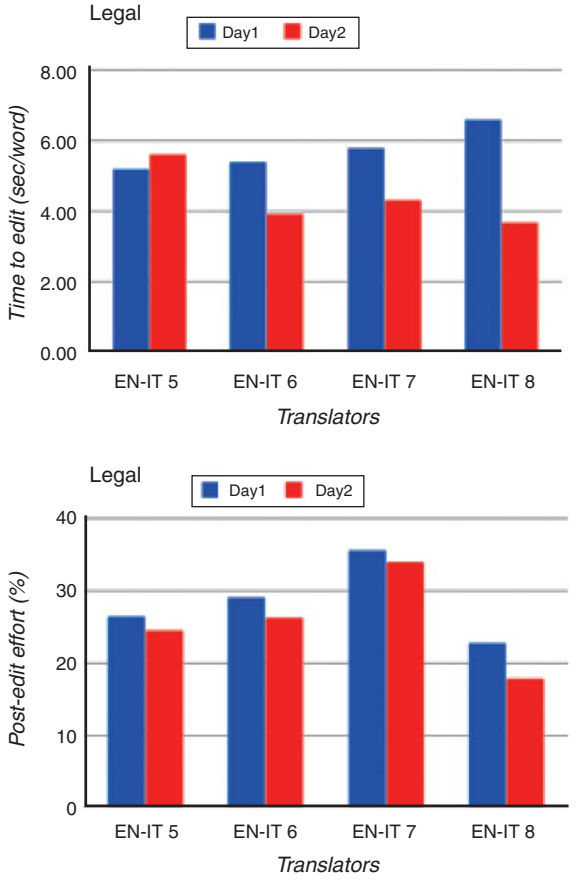

Fig. 2. Legal domain, English-Italian: time to edit and post-editing effort by four translators working with baseline MT system (Day 1) and self-tuned MT system (Day 2).

year, work on self-tuning MT focused on the investigation of data selection and model adaptation methods [7]. Results of the first field test have shown considerable improvements on the English-Italian direction (see Fig. 2 reporting results on the legal domain), both on post-editing effort (measured with HTER, *human translation error rate*), and time-to-edit (measured in seconds per word).

At this time, research on user-adaptive MT is exploring on-line learning based on generative [6] and discriminative methods [23, 39], while research on informative MT is currently focusing on quality estimation of MT output [37].

A major effort was also put in the development of the MateCat Tool. In particular, Translated srl developed the Web-based client application, while the research partners developed the MT server, compliant with the API Google Translate. The MT server implements an asynchronous multi-threaded version of Moses, so to manage independent requests by multiple MateCat clients. The GUI of the MateCat Tool was carefully designed in collaboration with professional translators in order to meet the standards of professional CAT tools, the code was optimized to provide fast interaction over the Web, the database was carefully designed to permit scalability and collaborative management of projects, and the collection and display of edit and time statistics was implemented in a way to allow reliable productivity measurements. The first version of the MateCat Tool has been publicly released in open source during 2012. A second updated version of the tool will be released by

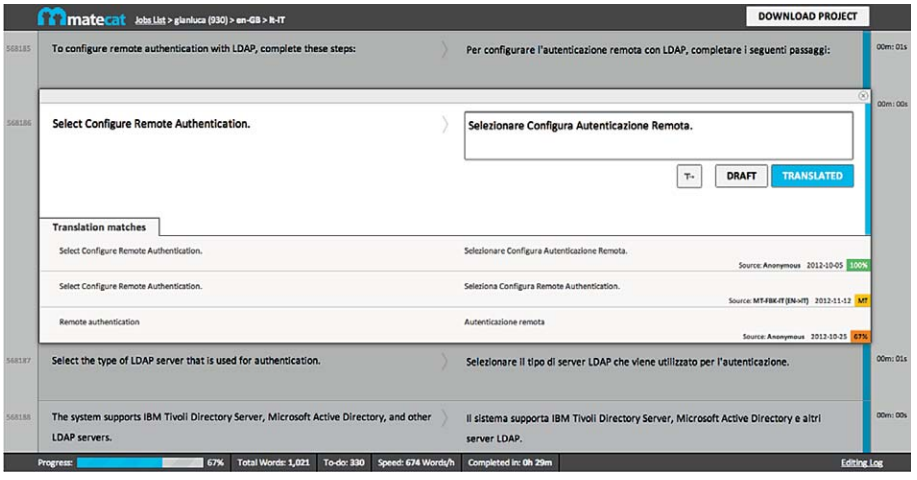

Fig. 3. Snapshot of the MateCat Tool. In the center is the pane of the current source segment (left,up), showing the edit window (right, up), and the translation suggestions (central, bottom).

the end of 2013. The tool is already employed by Translated on 15% of their internal translation projects, as well as tested by several international companies, both language service providers and IT companies. In this way, the project is collecting feedback from hundreds of translators, which besides helping us to improve the robustness of the tools is also influencing the way new MT functions will be integrated, in order to supply the best help to the final user.

6. Conclusion

In this paper we have presented a selection of relevant achievements in the HLT field where the role of Italian researchers has been crucial. Specifically, we have reported recent results in the following research directions: Distributional Semantics (i.e. the COMPOSES project), Multilingual Word Sense Disambiguation (the MultiJEDI project), Textual Semantic Inferences (the EXCITEMENT project), and Computer Assisted Translation (the MATECAT project). These projects, coordinated by Italian researchers, show both the capacity of the Italian community to compete at the highest level for very selective European funding, and the capacity to aggregate and to coordinate complex international initiatives on emerging research topics in Computational Linguistics.

Acknowledgments

The first author would like to thank the EXCITEMENT project, funded by the EU Commission under the grant FP7 ICT-287923. The second author would like to thank the ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES). The third author would like to thank the MATECAT project, funded by the EU Commission under the grant FP7 ICT-2011.4.2-287688. The last author gratefully acknowledges the support of the "MultiJEDI" ERC Starting Grant No. 259234.

References

[1] J. Aitchison, *Words in the Mind*. Blackwell, Malden, MA 1993.

[2] M. Baroni, R. Bernardi, N.-Q. Do and C.-C. Shan, Entailment above the word level in distributional semantics. In *Proceedings of EACL*, Avignon, France, pp. 23–32, 2012.

[3] M. Baroni, R. Bernardi and R. Zamparelli, Frege in space: A program for compositional distributional semantics. Submitted draft available from http://clic.cimec.unitn.it/composes, 2013.

[4] R. Basili and B. Magnini, Natural language processing in the web era, *Intelligenza Artificiale* **6**(02), 2012.

[5] R. Bernardi, G. Dinu, M. Marelli and M. Baroni, A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of ACL (Short Papers)*, Sofia, Bulgaria, In press, 2013.

[6] N. Bertoldi, M. Cettolo and M. Federico, Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Proceedings of the MT Summit XIV*, Nice, France, September 2013.

[7] N. Bertoldi, M. Cettolo and M. Federico, Project adaptation for mt-enhanced computer assisted translation. In *Proceedings of the MT Summit XIV*, Nice, France, September 2013.

[8] S. Bloehdorn and A. Moschitti, Combined syntactic and semantic kernels for text classification. In *Proceedings of ECIR*, Rome, Italy, pp. 307–318, 2007.

[9] N. Calzolari, B. Magnini, C. Soria and M. Speranza, *La Lingua Italiana nell'Era Digitale – The Italian Language in the Digital Age.* META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. Available online at http://www. meta-net.eu/whitepapers.

[10] I. Dagan, B. Dolan, B. Magnini and D. Roth, Recognizing textual entailment: Rational, evaluation and approaches, *Natural Language Engineering* **15**(4):i–xvii, 2009.

[11] I. Dagan and O. Glickman, Probabilistic textual entailment: Generic applied modeling of language variability. In *PASCAL workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France 2004.

[12] G. Dinu, N. The Pham, and M. Baroni, DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of the System Demonstrations of ACL*, Sofia, Bulgaria, In press, 2013.

[13] G. Dinu, N.T. Pham, and M. Baroni, A general framework for the estimation of distributional composition functions. In *Proceedings of ACL Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria, In press, 2013.

[14] P. Edmonds, *Designing a task for SENSEVAL-2*, Technical report, University of Brighton, U.K., 2000.

[15] K. Erk, Vector space models of word meaning and phrase meaning: A survey, *Language and Linguistics Compass* **6**(10):635–653, 2012.

[16] C. Fellbaum editor, *WordNet: An Electronic Database.* MIT Press, Cambridge, MA, 1998.

[17] T. Flati and R. Navigli, Spred: Large-scale harvesting of semantic predicates. In *Proc of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 2013.

[18] E. Grefenstette, G. Dinu, Y.-Z. Zhang, M. Sadrzadeh and M. Baroni, Multi-step regression learning for compositional distributional semantics. In *Proceedings of IWCS*, Potsdam, Germany, pp. 131–142, 2013.

[19] E.H. Hovy, R. Navigli, S.P. Ponzetto, Collaboratively built semi-structured content and artificial intelligence: The story so far, *Artificial Intelligence* **194**:2–27, 2013.

[20] A. Lazaridou, M. Marelli, R. Zamparelli and M. Baroni, Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of ACL*, Sofia, Bulgaria, pp. 1517–1526, 2013.

[21] A. Lazaridou, E.M. Vecchi and M. Baroni, Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In *Proceedings of EMNLP*, Seattle, WA, In press, 2013.

[22] B. Magnini, F. Cutugno, M. Falcone and E. Pianta, editors, Evaluation of Natural Language and Speech Tools for Italian, *International Workshop, EVALITA 2011*, Rome, Italy, January

24-25, 2012, Revised Selected Papers, volume 7689 of *Lecture Notes in Computer Science*, Springer, 2013.

[23] P. Mathur, M. Cettolo and M. Federico, Online learning approaches in computer assisted translation. In *Proceedings of the Eight Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August 2013.

[24] G. Miller and W. Charles, Contextual correlates of semantic similarity, *Language and Cognitive Processes* **6**(1):1–28, 1991.

[25] R. Montague, English as a formal language, *Linguaggi nella società e nella tecnica*, pp. 189–224, 1970.

[26] A. Moro, H. Li, S. Krause, F. Xu, R. Navigli and H. Uszkoreit, Semantic rule filtering for web-scale relation extraction. In *Proceedings of the 12th International Semantic Web Conference (ISWC 2013)*, Sydney, Australia 2013.

[27] A. Moro and R. Navigli, Integrating syntactic and semantic analysis into the open information extraction paradigm. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, Beijing, China, pp. 2148–2154, 2013.

[28] R. Navigli, Word Sense Disambiguation: A survey, *ACM Computing Surveys* **41**(2):1–69, 2009.

[29] R. Navigli, A quick tour of word sense disambiguation, induction and related approaches, *In Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, 115–129, 2012.

[30] R. Navigli, D. Jurgens and D. Vannella, Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, Atlanta, USA, pp. 222–231, 2013.

[31] R. Navigli and M. Lapata, An experimental study on graph connectivity for unsupervised Word Sense Disambiguation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(4):678–692, 2010.

[32] R. Navigli and S.P. Ponzetto, BabelNet: The automatic construction, evaluation and application of a wide coverage multilingual semantic network, *Artificial Intelligence* **193**:217–250, 2012.

[33] R. Navigli and S.P. Ponzetto, BabelRelate! a joint multilingual approach to computing semantic relatedness. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 2012.

[34] R. Navigli and 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 (EMNLP-CoNLL)*, 1399–1410, 2012.

[35] T.H. Ng, Getting serious aboutWord Sense Disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics:Why, What, and How?* Washington D.C, USA, pp. 1–7, 1997.

[36] N. The Pham, R. Bernardi, Y.-Z. Zhang and M. Baroni, Sentence paraphrase detection: When determiners and word order make the difference. In *Proceedings of the Towards a Formal Distributional Semantics Workshop at IWCS 2013*, Potsdam, Germany, pp. 21–29, 2013.

[37] M. Turchi, M. Negri and M. Federico, Coping with the subjectivity of human judgements in mt quality estimation. In *Proceedings of the Eight Workshop on Statistical Machine Translation*, Sofia, Bulgaria, 2013.

[38] P. Turney and P. Pantel, From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, **37**:141–188, 2010.

[39] K. Wäschle, P. Simianer, N. Bertoldi, S. Riezler, and M. Federico, Generative and discriminative methods for online adaptation in smt. In *Proceedings of the MT Summit XIV*, Nice, France, September 2013.