Natural Language Understanding: Instructions for (Present and Future) Use

Roberto Navigli
Department of Computer Science
Sapienza University of Rome
Viale Regina Elena 295, Rome, Italy
navigli@di.uniroma1.it

Abstract

In this paper I look at Natural Language Understanding, an area of Natural Language Processing aimed at making sense of text, through the lens of a visionary future: what do we expect a machine should be able to understand? and what are the key dimensions that require the attention of researchers to make this dream come true?

1 Introduction

Natural Language Processing (NLP) is a challenging field of Artificial Intelligence which is aimed at addressing the issue of automatically processing human language, called natural language, in written form. This is to be achieved by way of the automatic analysis, understanding and generation of language. While all these tasks are difficult for a machine to perform, Natural Language Understanding (NLU) – which involves a semantic (and a pragmatic) level – is particularly challenging owing to the pervasive ambiguity of language and the subtly different perceptions humans have of word, phrase and sentence meanings. NLU aims to make sense of language by enabling computers to read and comprehend text. The key question is, therefore, how to derive meaning from natural language by overcoming its inherent complexities. NLU, in turn, is expected to enable one of the longstanding goals of AI, that is, machine reading [Etzioni *et al.*, 2007]. And what would a machine "learn" as a result of reading and understanding text, and storing a semantic representation of it? Well, new scenarios would open up whereby the machine became enabled to analyze, aggregate and reason on huge amounts of information and perform tasks that would be precluded to humans, simply because of the scale and time involved.

To make this dream come true, however, we need computers capable of inputting text and outputting semantic representations, something that requires a full-fledged NLP architecture in multiple languages. But while recent progress has made it possible to perform text processing in dozens of languages up to the syntactic level, semantic analysis is still a big challenge, especially if we want the machine to be able to comprehend text in arbitrary languages. Recent and ongoing work in this direction has been aiming to develop algorithms that can process open text and produce structured semantic representations which, ideally, are independent of the language they were obtained from and of the way they were expressed in that language. Before delving into the various dimensions and "modes" of use, I will briefly review some history of NLU.

2 Before You Get Started (a Brief History)

The story of AI has its origin in "an ancient wish to forge the gods" [McCorduck, 2004]. But can we communicate with our creatures? In history, mythological gods used to talk to humans and, indeed, AI and NLP are inextricably bound to each other: when thinking about what might provide proof of intelligence, as early as 1950 it occurred to Alan Turing that language was the natural answer [Turing, 1950]. And in fact we humans often determine the intelligence of our peers based on verbal communication.

Ironically, the earliest form of (only seemingly) intelligent NLU was simply a reflection of our expectations: the Eliza experiment, after the initial illusion of understanding it gave, hinted at the fact that processing language just as mere strings is not what humans do. Text implies knowledge of concepts and of the real world, it requires further reasoning, it arouses emotions. The dream of formalizing, encoding and later exploiting such knowledge was nurtured and advanced with groundbreaking work on frames [Minsky, 1975] and large projects that are still alive and thriving today, such as WordNet [Miller *et al.*, 1990], FrameNet [Baker *et al.*, 1998] and BabelNet [Navigli and Ponzetto, 2012].

The surge of interest in statistical techniques for NLP and supervised machine learning in the 1990s led to mainstream approaches centered around probabilistic frameworks and high-performance classifiers that could focus on a specific task in processing text, such as tokenization, part-of-speech tagging and syntactic parsing. However, the most impressive NLP application made available to the general public was statistical machine translation, which, after a first series of word-based approaches put forward by IBM, progressed in the 2000s to the key idea of translating phrases and not just words, and then recombining the translated phrases in the target language [Koehn *et al.*, 2003]. More recently, the advent of deep learning has again revolutionized the way machine translation is performed, attaining even better performances with results that are surprisingly good. The question remains, however, as to whether the system really understands the text being translated, or, more realistically, whether it merely mimics the Chinese room metaphor [Searle, 1980].

While neural networks have brought important improvements in virtually all areas of Natural Language Processing, it is the popularization of word embeddings that has changed the landscape of most lexical-semantic tasks, that is, those tasks which deal with the meaning of language. At the core of this issue lies the lexical ambiguity of language, a problem addressed by a key field in computational lexical semantics, namely Word Sense Disambiguation [Navigli, 2009, WSD]. If we look back at the 1990s and the early 2000s, WSD was performed mainly in the English language, with a fine-grained inventory of senses and disappointing results which struggled to surpass 65% accuracy in an all-words disambiguation setting. Today, thanks to the introduction of a multilingual knowledge resource like BabelNet [Navigli and Ponzetto, 2012], we have word sense disambiguation systems that can scale to hundreds of languages while at the same time also performing the entity linking task [Moro *et al.*, 2014]. But while the most recent LSTM-based approaches achieve above 70% accuracy on difficult datasets [Yuan *et al.*, 2016; Raganato *et al.*, 2017], several limits still have to be overcome, which I will discuss in the remainder of this paper.

As we move on from words to sentences and move towards enabling machine comprehension of text, it is important to focus on predicates and the expected roles that their arguments can potentially fill. This task, called Semantic Role Labeling (SRL) [Gildea and Jurafsky, 2002], consists of automatically detecting the semantic arguments of a given predicate, typically the focus verb of a sentence, and classifying such arguments into their specific roles. First, although frameset repositories such as PropBank and FrameNet paved the way for the task, a number of issues inherently affect them, among which we cite: partial coverage of the lexicon, low coverage of languages, paucity of large-scale training data. These issues are even more important for the more general task of semantic parsing, whose objective is to map sentences to formal representations of their meaning. On the positive side, semantic parsing goes deeper in understanding language, and is therefore the task that, more than any other, would seem to hold the potential to achieve the ambitious objective of machine reading.

3 Modes of Use

3.1 The Revolution of Deep Learning

The last few years have seen a revolution in the way NLP is implemented and innovated thanks to the (re)introduction of neural networks and deep learning. The key improvements over pre-neural approaches are undoubtedly the considerable reduction of data sparsity and the compactness of the lexical representations. These come, however, at the cost of flattening information and, at least initially, conflating the meanings of ambiguous words into a single vector representation. More importantly, the biggest challenge of neural approaches is their accountability in the future, i.e., the ability to explain their outputs in a way that makes it possible to apply remedies. While this is an obvious issue for driverless cars, it is also important that an intelligent system should be able to explain the process followed for understanding text, especially if a decision has to be taken (e.g., in booking a restaurant or fixing an appointment by interacting with a vocal assistant).

3.2 Explicit vs. Implicit

This brings us to the question of whether meaning representations should be implicit or explicit, or maybe both.

The implicit approach. Much of the work in lexical semantics is currently centered around learning and exploiting embeddings of words, contexts and sentences, something which comes naturally from the training process of neural networks, from the simplest feed-forward networks to Long-Short Term Memory (LSTM) architectures. In this scenario, we have the conflation of senses of the same word into a single latent word representation. However, this conflation is somewhat compensated for by training on large numbers of word sequences, for instance by using context [Melamud *et al.*, 2016] or sentence vectors [Yuan *et al.*, 2016] to perform the NLU task.

The explicit approach. A different strand of work in semantic vector representation stems from the use of lexical-semantic knowledge resources to link vectors with explicit concept entries, such as the word senses and synsets available in the WordNet computational lexicon [Miller *et al.*, 1990]. The big advantage of this approach lies in its capability of discriminating between the senses of the various words and, more importantly, to give adequate coverage to infrequent meanings which, nevertheless, might play a crucial role within a given domain (e.g., the predominant meaning of *bus* is clearly the vehicle, but in a computer science text the dominant usage would surely shift towards the hardware bus meaning), a point which I cover in more detail in Section 3.6. A second big advantage of the explicit approach is that scaling to multiple languages is made easy (this point is discussed in more detail in Section 3.8).

Explicit and implicit together. Recent developments have shown that the implicit and explicit approaches can live together: vector representations for senses and synsets can be embedded in the same vector space as word representations. Among these, we cite the SensEmbed sense representations that are obtained by inputting sense-annotated text to learn word and sense embeddings in the same space [Iacobacci *et al.*, 2015]; AutoExtend, a lemma-sense-synset constraint-based approach [Rothe and Schütze, 2015]; the NASARI embedded representations of synsets [Camacho-Collados *et al.*, 2016], where a synset embedding is given by the weighted average of the embeddings of the most relevant words which are used in a subcorpus to describe the concept; DeConf [Pilehvar and Collier, 2016], where sense vectors are created with the constraints that the representations of senses of the same word should be close to their word representation while also being close to sense-specific biasing words; SW2V [Mancini *et al.*, 2017], where the CBOW word2vec architecture is extended to include input and output senses obtained automatically via a knowledge-based approach (cf. Section 3.7).

3.3 Word Senses: Which Representation?

Interestingly, the choice between an implicit or an explicit representation mostly concerns the lexical level. As we focus on any of the key lexical semantic tasks, from WSD to SRL and semantic parsing, we see that – independently of the input, internal and output representations a neural network architecture processes – systems have to assign explicit labels from an existing inventory, unless we resort to an unsupervised approach (word sense induction [Navigli, 2009] or unsupervised SRL, cf. Section 3.7). We therefore now provide a review of the main inventory options for word sense representation.

WordNet [Miller *et al.*, 1990]. The oldest wide-coverage computational lexicon of English, a project started in 1985, is still very actively used, thanks to its organization of lexical knowledge in synsets (synonym sets) based on psycholinguistic principles. Much has been written on the limits of the fine granularity of WordNet and several, non-conclusive proposals have been put forward which make sense distinctions coarser [Navigli, 2006; Snow *et al.*, 2007; Hovy *et al.*, 2006]. Today, with the possibility to mix implicit and explicit approaches, the issue is partially mitigated because fine-grained senses of a word will be close in the semantic vector space. However, gold-standard datasets from the Senseval and SemEval competitions are still annotated with WordNet senses, which forces WSD systems to work with that level of granularity. The only exceptions are the two coarse-grained datasets created at SemEval-2007: the current state of the art is in the range of 72-74% for fine-grained WSD [Raganato *et al.*, 2017; Yuan *et al.*, 2016] and 84% for coarse-grained WSD [Moro *et al.*, 2014].

Oxford Dictionary of English and proprietary resources. The Oxford Dictionary of English (ODE) and other proprietary resources like the New Oxford American Dictionary (NOAD) have been used in some work to tackle some of the limits of WordNet [Navigli, 2006; Yuan *et al.*, 2016]. ODE and NOAD, in fact, group senses into homonyms, core senses and subsenses, which makes it easy to select the most suitable granularity based on the task at hand. However, this comes at the cost of preventing the research community from using this data widely and, because of this, lacking training data. Indeed, the biggest merit of WordNet, which has led to its widespread adoption, is its free availability. An endeavour in the direction of making such precious resources contribute to the research in NLU is offered by the ELEXIS project, which aims to create a European Lexicographic Infrastructure thanks to which open and proprietary resources will live in a common, interoperable space.

BabelNet [Navigli and Ponzetto, 2012]. The most recent large-scale work aimed at providing an inventory of meanings, BabelNet[1] brings together the lexical-semantic knowledge available in WordNet, as well as wordnets in other languages [Bond and Foster, 2013, Open Multilingual Wordnet] and the most widespread crowdsourced resources such as Wikipedia, Wiktionary and, more recently, Wikidata. BabelNet scales the WordNet synset model to include lexicalizations in multiple languages: a multilingual synset is a collection of senses coming from the various resources interlinked in BabelNet. As a result, WSD in arbitrary languages is enabled [Moro *et al.*, 2014] and training and test data can be annotated with BabelNet synsets which implicitly provide annotations to the various resources from which each synset is made up.

3.4 Semantic Roles: Which Representation?

We now move from word senses to semantic roles, which express the abstract roles taken by the predicate arguments in a given event and define the possible classes used for SRL.

FrameNet. Based on a theory of meaning called Frame Semantics [Fillmore, 1982], FrameNet [Baker *et al.*, 1998] is a lexical database of English at the core of which are semantic frames, i.e., manually-curated conceptual structures that describe events, relations or entities, and the participants that are involved in them. In contrast to the other mainstream resources for SRL, frames encode an abstract notion of event which can be expressed with different verbs and nouns. The dataset comes with thousands of sentences annotated with frames. Coverage of the English FrameNet is limited to 13640 lexical units, among which 5200 verbal units, around half of which are marked as finished.

VerbNet. A different approach is taken by VerbNet [Kipper *et al.*, 2000], which provides a large lexicon of English verbs, organized hierarchically with mappings to other resources like WordNet and FrameNet. Verbs are arranged into Levin's verb classes [Levin, 1993], so as to preserve syntactic and semantic coherence within each class. Because it provides additional information, such as thematic roles and hierarchically-organized selectional restrictions for each class, VerbNet is the most complex of the SRL resources. However, it is not widely adopted for the SRL task.

PropBank. A more recent resource is PropBank [Palmer *et al.*, 2005], a text collection manually annotated with verbal propositions and their arguments. In contrast to FrameNet, PropBank is centered around the linguistic notion of verb, rather than the more abstract notion of frames. Interestingly, PropBank was originally designed with the idea of SRL in mind, and this is one reason why it remains closer to the syntactic level. All of the English CoNLL datasets for SRL are annotated with PropBank framesets.

There are two main limitations affecting semantic roles. First, although frameset repositories paved the way for the SRL task, two key issues inherently affect them. First, their coverage of English verbs is limited to some thousand items, typically below 30% of the overall set of verb meanings (e.g., compared to WordNet). Therefore SRL systems have to face the additional difficulty of not knowing whether a low-probability labeling is due, or not, to poor coverage of that

[1] http://babelnet.org

verb and its roles in the reference resource. A second, crucial issue, related to the above, is that these resources cannot easily scale to a multitude of languages, nor are they interconnected across languages. Only sporadic research has focused on providing similar resources in other languages, with even lower coverage than English, unfortunately.

3.5 Semantic Parses: Which Representation?

As we move from SRL to semantic parsing, things get more complex. Determining which representation should be used to describe a sentence's semantics involves choosing among the many existing representations, ranging from the Discourse Representation Theory used in the Groningen Meaning Bank [Basile *et al.*, 2012] to the Universal Conceptual Cognitive Annotation [Abend and Rappoport, 2013, UCCA], and the Abstract Meaning Representation (AMR) formalism [Banarescu *et al.*, 2013]. This last representation, which adopts PropBank for its predicates, is gaining increasing attention from the NLP community, with several recent approaches to semantic parsing producing AMR representations [Foland and Martin, 2017; Damonte *et al.*, 2017] and a relatively large corpus, the AMR Bank[2], containing some tens of thousands of sentences annotated with AMR graphs.

3.6 Word Sense and Semantic Role Skewness

We have known that words follow a Zipfian distribution, i.e., that the word types that occur within a text are distributed according to a power law [Zipf, 1935], for more than a century. More recent studies also tell us that word senses are distributed according to some power law, too [Kilgarriff, 2004]. This implies that the most frequent sense of an ambiguous word will occur on average two thirds of the time: this results, first, in a hard-to-beat baseline calculated by counting senses from training data, and, second, in making it hard for a WSD system to discriminate between more than two, sometimes three, most frequent senses of a word, which would already account for 94% of the occurrences (in SemCor). This opens up the question as to whether it even makes sense to aim for disambiguating very infrequent senses, something that might be done better with knowledge-based approaches to WSD.

As regards SRL, this phenomenon is even more amplified, as the two predominant proto-roles, i.e., proto-agent and proto-patient (respectively, ARG0 and ARG1 in PropBank) have the widest prevalence, so the doubt is whether current supervised SRL systems are really grasping the semantics of sentences, or are just classifying the two predominant classes correctly, while neglecting the other infrequent roles.

3.7 To Supervise or Not to Supervise

Supervised vs. Unsupervised. The supervised paradigm, where we train a machine learning system with linguistic items tagged with the most suitable classes, works well in the presence of an adequate amount of annotated data. Unfortunately, however, while we could come up with 17 universal part-of-speech classes for each of which it is easy to provide tens of thousands of example occurrences, an equivalent effort in computational lexical semantics is rendered more complex by the lack of agreement as to which inventories and formalisms to use (cf. Section 3.3-3.5) and the problems associated with each of them. A way out is to avoid predefined classes, which makes the problem unsupervised and paves the way to using huge amounts of raw text. In all the three tasks (WSD, SRL and semantic parsing) we have seen unsupervised approaches: graph-based, probabilistic and neural Word Sense Induction, unsupervised SRL, and semantic parsing without training data. However, the unsupervised task performs poorly and is recommended only for low-resourced languages. The task, indeed, would be hard, if not impossible, to perform even for humans if they were to have no knowledge of the language and of the real world (which reminds us again of the Chinese room argument).

Supervision vs. Knowledge. A more engaging challenge is whether to use supervision, in the sense of traditional training data for a machine learning model, or knowledge, which is a paradigm that is fairly specific to computational lexical semantics and NLU. The knowledge-based paradigm differs from the supervised paradigm in the way the classifier is informed: it exploits knowledge from external resources, typically provided in structured form, like WordNet, BabelNet or another lexical-semantic resource (cf. Section 3.3), to perform the task. Importantly, the resource is not task-specific, and therefore provides general information that can be utilized in other tasks, too. Examples of current knowledge-based approaches in WSD are those based on Personalized PageRank to perform the task [Agirre *et al.*, 2014], as well as densest graph approximation algorithms to choose from the sense candidates in a knowledge graph of the context [Moro *et al.*, 2014].

Compared to supervised approaches, knowledge-based ones have the big advantage of coming with wide coverage, but without having to draw on large, expensive annotation jobs. Adopting resources like BabelNet enables multilinguality at virtually no additional effort (except for the ability to preprocess text, which includes tokenization, part-of-speech tagging and, sometimes, lemmatization). This claim is substantiated by the fact that the current state-of-the-art performance in WSD on non-English languages has been achieved by knowledge-based systems, whereas on English the performance of knowledge-based systems is competitive [Moro *et al.*, 2014], unless massive amounts of training data are provided for each word [Yuan *et al.*, 2016]. In SRL, where semantic roles can potentially be learned independently of the verb, sparsity is less of an issue, therefore the state of the art is supervised. Finally, due to its complexity and training data scarcity, semantic parsing is also sometimes performed in a knowledge-, graph-based fashion with good performance.

3.8 Scalability and Multilinguality

In order to work on open text, NLU systems need to scale in a number of respects. First, can the system work on the entire lexicon? Given the current trend of word embeddings, scaling to the whole lexicon is realistic thanks to the reduction of data sparsity. Second, can the system work in multiple languages? This is again made possible by the most recent developments on bilingual and multilingual embeddings in a

[2] https://amr.isi.edu/

shared vector space [Conneau *et al.*, 2017]. In this case, however, even if training in one language and testing in another language is a challenging option, e.g., leading to competitive results in WSD [Raganato *et al.*, 2017], getting a level playing field across languages is still an open issue (see Section 3.10 below for recent options). Furthermore, while scalability seems at hand for words, it is not obvious that what is learned by a supervised system for certain words at the level of either word senses or semantic roles can easily scale to other senses or roles in the same language. The problem is mitigated if we embrace the knowledge-based paradigm, as structured knowledge and the wide coverage of the resources adopted make large-scale processing possible, also across languages if such a resource is multilingual by design.

3.9 Universality and Language Independence

The universal POS and syntactic dependency tagsets[3] are a clear example of how work done independently in multiple languages (and apparently without clear connections) can be brought together in a unified framework. Semantics, however, is more difficult: while a verb is a verb (even though I expect some linguists to disagree on this), establishing the senses of a given verb is far from trivial (and there is no obvious reason for choosing among two reasonable sets of senses if not for granularity issues, cf. Section 3.3), not to mention agreeing on the set of predicates and semantic roles to use in SRL and semantic parsing. Also, 17 universal classes were identified for POS tagging, while more than a hundred thousand meanings make up a WordNet-like computational lexicon (or millions, in the case of an encyclopedic dictionary like BabelNet). However, the die is cast: it is unavoidable that future research in this direction will progress towards universal senses and roles, so as to enable not only the scalability of systems, but also – and maybe more importantly – the independence both of their language and of the representations they output.

3.10 The Knowledge Acquisition Bottleneck

The scarcity of semantically-annotated data is called the knowledge acquisition bottleneck. As mentioned above, as we move to hundreds of thousands of classes, expecting an adequate amount of manually tagged data is unrealistic. This issue has affected the field of WSD for decades, with the largest manually-curated dataset dating back to 25 years ago [Miller *et al.*, 1993]. While important efforts have been undertaken to produce more data, including the Manually-Annotated Sub-Corpus, OntoNotes, and some CoNLL datasets, we are still in search of bigger datasets, above all for non-English languages. A current direction is to produce such annotated data semi-automatically, e.g., by exploiting bilingual translations [Taghipour and Ng, 2015] or by applying knowledge-based systems which can perform the task on millions of sentences with very high precision and low recall, so as to select those items which can be reliably used later for training a supervised system. For instance, Train-O-Matic[4] [Pasini and Navigli, 2017] has demonstrated high WSD performance (sometimes even better than the state of the art) without any manual annotation of text. As we move to SRL, the challenge might not be harder if the annotation is predicate-independent – e.g., if we disregard the predicate specificity of PropBank arguments. As regards semantic parsing, the challenge is, instead, more difficult, as we need a considerably higher number of sentences annotated with their semantic structures to perform the task on open text.

4 Conclusion

NLP is not just an application of machine learning [Manning, 2015], and having discussed the many peculiarities and challenges of NLU here, this appears particularly obvious. As a result of their interactions, words create meanings and such meanings cannot easily be grasped with crisp classes, unless these classes are structured in a way that enables inference. A fundamental question is whether we need explicit lexical semantics at all in order to perform NLU: cannot we just understand text by comparing and transforming its latent representations to those of other texts, without making the effort of identifying and associating explicit semantics? Potentially yes, but would it be better? Posterity will judge.

Acknowledgments

The author wishes to thank Michele Bevilacqua, Jim McManus and Federico Scozzafava for their invaluable help. This work was partially funded by Babelscape, the MOUSSE ERC Consolidator Grant (contract no. 208488) and the ELEXIS projects (contract no. 731015) under the European Union's Horizon 2020 research and innovation programme.

References

[Abend and Rappoport, 2013] Omri Abend and Ari Rappoport. Universal conceptual cognitive annotation (UCCA). In *Proc. of ACL*, pages 228–238, 2013.

[Agirre *et al.*, 2014] Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84, 2014.

[Baker *et al.*, 1998] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley Framenet project. In *Proc. of ACL*, pages 86–90, 1998.

[Banarescu *et al.*, 2013] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proc. of 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 2013.

[Basile *et al.*, 2012] Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. Negation detection with discourse representation structures. In **SEM*, 2012.

[Bond and Foster, 2013] Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In *Proc. of ACL*, pages 1352–1362, 2013.

[Camacho-Collados *et al.*, 2016] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli.

[3] http://universaldependencies.org/

[4] http://trainomatic.org

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities. *Artificial Intelligence*, 240:36–64, 2016.

[Conneau *et al.*, 2017] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[Damonte *et al.*, 2017] Marco Damonte, Shay B. Cohen, and Giorgio Satta. An incremental parser for abstract meaning representation. In *Proc. of EACL*, pages 536–546, 2017.

[Etzioni *et al.*, 2007] Oren Etzioni, Michele Banko, and Michael J. Cafarella. Machine reading. In *2007 AAAI Spring Symposium*, pages 1–5, 2007.

[Fillmore, 1982] Charles Fillmore. Frame semantics. *Linguistics in the morning calm*, pages 111–137, 1982.

[Foland and Martin, 2017] William Foland and James H. Martin. Abstract meaning representation parsing using LSTM recurrent neural networks. In *ACL*, 2017.

[Gildea and Jurafsky, 2002] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[Hovy *et al.*, 2006] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: The 90% solution. In *HLT-NAACL*, 2006.

[Iacobacci *et al.*, 2015] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Sensembed: Learning sense embeddings for word and relational similarity. In *Proc. of ACL*, pages 95–105, 2015.

[Kilgarriff, 2004] Adam Kilgarriff. How dominant is the commonest sense of a word? In *Text, Speech and Dialogue*, pages 103–111. Springer, 2004.

[Kipper *et al.*, 2000] Karin Kipper, Hoa Trang Dang, and Martha Stone Palmer. Class-based construction of a verb lexicon. In *Proc. of AAAI*, pages 691–696, 2000.

[Koehn *et al.*, 2003] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proc. of NAACL*, pages 48–54, 2003.

[Levin, 1993] Beth Levin. *English verb classes and alternations: A preliminary investigation*. U.Chicago press, 1993.

[Mancini *et al.*, 2017] Massimiliano Mancini, José Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. Embedding words and senses together via joint knowledge-enhanced training. In *CoNLL*, 2017.

[Manning, 2015] Christopher D. Manning. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707, 2015.

[McCorduck, 2004] Pamela McCorduck. *Machines Who Think*. A. K. Peters, 2004.

[Melamud *et al.*, 2016] Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning generic context embedding with bidirectional LSTM. In *CoNLL*, 2016.

[Miller *et al.*, 1990] George A. Miller, R.T. Beckwith, C. D. Fellbaum, D. Gross, and K. Miller. WordNet: an online lexical database. *Int J. Lexicography*, 3(4):235–244, 1990.

[Miller *et al.*, 1993] George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. A semantic concordance. In *Proc. of 3rd Workshop on HLT*, pages 303–308, 1993.

[Minsky, 1975] Marvin Minsky. A framework for representing knowledge. AI memo, MIT, 1975.

[Moro *et al.*, 2014] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244, 2014.

[Navigli and Ponzetto, 2012] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[Navigli, 2006] Roberto Navigli. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proc. of ACL 2006*, pages 105–112, 2006.

[Navigli, 2009] Roberto Navigli. Word Sense Disambiguation: a survey. *ACM Comput. Surveys*, 41(2):1–69, 2009.

[Palmer *et al.*, 2005] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

[Pasini and Navigli, 2017] Tommaso Pasini and Roberto Navigli. Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages without Manual Training Data. In *EMNLP*, pages 78–88, 2017.

[Pilehvar and Collier, 2016] Mohammad Taher Pilehvar and Nigel Collier. De-conflated semantic representations. In *Proc. of EMNLP*, pages 1680–1690, 2016.

[Raganato *et al.*, 2017] Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. Neural Sequence Learning Models for Word Sense Disambiguation. In *Proc. of EMNLP 2017*, pages 1156–1167, 2017.

[Rothe and Schütze, 2015] Sascha Rothe and Hinrich Schütze. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *ACL*, 2015.

[Searle, 1980] John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–424, 1980.

[Snow *et al.*, 2007] Rion Snow, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. Learning to merge word senses. In *Proc. of EMNLP-CoNLL 2007*, pages 1005–1014, 2007.

[Taghipour and Ng, 2015] Kaveh Taghipour and Hwee Tou Ng. One million sense-tagged instances for word sense disambiguation and induction. In *CoNLL*, 2015.

[Turing, 1950] Alan M. Turing. Computing machinery and intelligence. *Mind*, 54:443–460, 1950.

[Yuan *et al.*, 2016] Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. Semi-supervised word sense disambiguation with neural models. In *Proc. of COLING*, pages 1374—1385, 2016.

[Zipf, 1935] George Kingsley Zipf. *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin, 1935.