How Target Sense Verification improves domain-specific and enterprise disambiguation settings

Anna Breit and Artem Revenko

Introduction

Polysemous words are inherent to natural language. While we as humans usually have little trouble processing them, they can have a significant impact on downstream Natural Language processing (NLP) tasks.

For example, classical static word embeddings such as Word2Vec (Mikolov et al. 2013) create distributed word representations in vector space, such that similar words are close together. Without further processing of polysemous words, one vector will be produced for each surface form, instead of for each meaning. This conflation deficiency yields to skewed semantic vector representations, as the word vectors are optimised for all senses simultaneously. In consequence, semantically unrelated words, which are similar to a particular sense of the ambiguous word, are pulled together in the vector space (Camacho-Collados and Pilehvar 2018). For example *hint*, *clue* and *evidence* will end up close to *cadmium*, *mercury* and *silver*, since they share *lead* as a semantically similar word.

Therefore, to minimise the negative effects of ambiguous words on downstream NLP tasks, it is essential to correctly disambiguate them.

In this article, we give a brief introduction on existing disambiguation task formulations and highlight their shortcomings, especially in the light of modern domain-specific and enterprise settings. We will introduce Target Sense Verification (TSV), a reformulation of the disambiguation task, which aims to overcome these drawbacks. Finally, we present Word in Context Target Sense Verification, a novel benchmark dataset based on the TSV formulation, which enables investigating both the generalisation capabilities of models, as well as their ability to transfer the knowledge gained on the general domain into specific domain settings.

Existing disambiguation task formulations

A long-standing task formulation is Word Sense Disambiguation, where the goal is to find out which sense of a word is used in a given

Anna Breit joined Semantic Web Company in 2019 in the role of Research Engineer and participates in the European research project Prêt-à-LLOD as well as in the Austrian research project OBARIS. She holds a master's degree in medical onformatics, with a focus on Neuroinformatics, from the Medical University in Vienna. Her research focuses on knowledge graphs and natural language processing, with a particular interest in representation learning, and she deals also with questions of evaluation design and setup.
anna.breit@semantic-web.com

sentence by comparing the word in context with all available senses in a provided sense inventory. The most prominent sense inventory used for this task in NLP context is Princeton WordNet (Fellbaum 1998), a large lexical database containing hierarchically organised concepts (*Synsets*). Although this inventory covers a wide range of senses, WordNet focuses on representing general-domain concepts, while the coverage of domain-specific terms and named entities is limited. Furthermore, as general sense inventories are complex to maintain, they often lag behind in being up to date, yielding to the absence of novel terms and term usages. Even though efforts are being made to improve the topicality of wordnets by introducing open-source versions, the absence of some senses still seems inevitable.

On the other hand, crowd-sourced encyclopaedias, such as Wikipedia, aim at describing proper nouns referring to real-world entities. However, in fact, Wikipedia also contains pages for abstract general concepts and notable domain-specific entities. Machine-readable forms of Wikipedia, such as the open knowledge graph DBpedia (Auer et al. 2007), thus offer an alternative to traditional sense inventories for performing disambiguation, where this task is referred to as Entity Disambiguation. The crowd-sourcing nature of these open knowledge graphs positively contributes to the topicality and coverage of contained concepts.

Suitability for modern enterprises and domain-specific disambiguation settings

Although the presented task formulations have a long tradition in NLP and drove a wide variety of research efforts, they also come with disadvantages, especially in modern enterprise and domain-specific disambiguation settings. The drawbacks mainly originate from the focus on finding the most suitable sense or entry in a sense inventory or knowledge graph, ultimately requiring systems to model the senses according to the underlying knowledge structure.

This reduces the flexibility of the models that try to solve the presented task. More specifically, the models are dependent on the knowledge structure, as changes in the structure require changes to the model.

Furthermore, this formulation assumes the availability of all senses whereas, as mentioned earlier, general sense inventories are often not up-to-date, and their coverage of domain-specific terms is limited. Even though open knowledge graphs do contain notable domain-specific entities, they can only cover a specific domain to a certain extent. Open domain-specific sense inventories and knowledge

Artem Revenko is Director of Research at Semantic Web Company. He holds PhD titles in mathematics from Lomonosov Moscow State University and in computer sciences from TU Dresden and TU Vienna, in the framework of European PhD Program in Computational Logic. He participates in several European and Austrian research projects, including Prêt-à-LLOD, Lynx, WATEREYE and OBARIS, and his research interests lie in the intersection of machine learning (primarily natural language processing) and formalizing knowledge.
artem.revenko@semantic-web.com

graphs, on the other hand, are rare for most domains and, in many cases, incomplete. As a result, the creation of enterprise knowledge graphs takes a notable amount of effort, yielding missing senses.

Last but not least, when modelling entire sense inventories, all senses have to be taken into account while, in reality, only a small number of senses will be interesting in a specific use case.

The impact of these drawbacks is best illustrated in a real-world example. Assuming the target domain 'information technology' (IT) and the collection of information on the current IT landscape as a goal, the following context needs to be disambiguated to evaluate its relevance:

> *Terno Schwab has been Chairman of the Executive Board of Alphabet since January 2017.*

Incorporating a general sense inventory into the disambiguation system of 'Alphabet' would add common senses such as the "set of letters used to write a language", or the "simplest rudiments of a subject", while open knowledge graphs could contribute prominent entities named Alphabet, including Alphabet Inc., the parent company of Google. The technology company Alphabet would also be present in a considered domain-specific knowledge base, which could furthermore extend the list of senses with the python library *alphabet*. However, even when incorporating a general sense inventory, an open knowledge graph, and a technology-specific sense inventory, the actual target sense of this context – i.e., Alphabet Austria Fuhrparkmanagement GmbH, an Austrian fleet management company – would still be missing, which makes the annotation of the correct sense impossible.

For these reasons, the current disambiguation task formulations and existing benchmarks that build on top of them are not fully able to evaluate the suitability of disambiguation systems in realistic domain-specific and/or enterprise settings.

Target Sense Verification

From the aforementioned shortcomings of disambiguation task formulations, two aspects whose improvement would lead to enhanced suitability for enterprise and domain-specific settings can be identified.

First, the dependency on sense inventories forms restrictions to the disambiguation system's flexibility. Therefore, the requirements that are implied by the incorporated knowledge bases should be reduced to a minimum. One way to achieve this is by determining a common

minimal template that defines a sense, i.e., specifying sense indicators such as definition and hypernyms.

Second, as it is not realistic to have all senses available in domain-specific and enterprise settings, it is necessary to minimise the number of senses that are required to be present. Ideally, we would only need to know a single sense.

Based on these investigations, we introduce a reformulation of the disambiguation task: Target Sense Verification. TSV formulates the disambiguation of a word as a binary classification task where the equivalence of the intended sense of a word in context and a single given sense are evaluated. Therefore, instead of comparing a target word against all possible senses, it is only verified against one target sense.

For instance, in the example above, the system would need to decide whether the sentence refers to Alphabet Inc., the technology company, by being provided with sense indicators for Alphabet Inc. only (e.g. the hypernym *technology company* and definition "Alphabet Inc. is an American multinational conglomerate headquartered in Mountain View, California and the parent company of Google").

This new formulation comes with the advantage that existing enterprise and domain-specific senses can be easily used as target senses for TSV, regardless of their current representation in the use case environment. As sense indicators aim to be as generic as possible, they can be easily generated from all kinds of knowledge representations. Moreover, when creating a new resource for a domain-specific or enterprise use case, there is no need to take out-of-domain senses into account. Finally, pre-training and domain adaptation can be more easily exploited, as the requirements for the domain-specific use case and resources are minimal.

WiC-TSV: The benchmark

The Word in Context Target Sense Verification (WiC-TSV) benchmark (Breit et al. 2021) was designed on top of the TSV task formulation, and thus pursues the goal of evaluating the ability of a disambiguation system to verify the target sense of a word in a context without the usage of an external sense inventory or knowledge graph, i.e., without knowing all possible senses of the target word.

Formally, each instance in the dataset consists of a target word w, a context c containing w, and the corresponding target sense s represented by its definition and hypernyms. The task aims to determine whether the intended sense of the word w used in the context c matches the target sense s.

The totality of 3832 available instances created for this benchmark was split into training, development, and test sets with a ratio of 56:10:34, which allows for a sophisticated analysis of the generalisation capabilities of tested systems, while still providing an appropriately-sized training set.

To test the disambiguation system's capability to classify usages of previously unseen words, care was taken to ensure that the overlap of target senses in the training and the test was particularly small (below 3%).

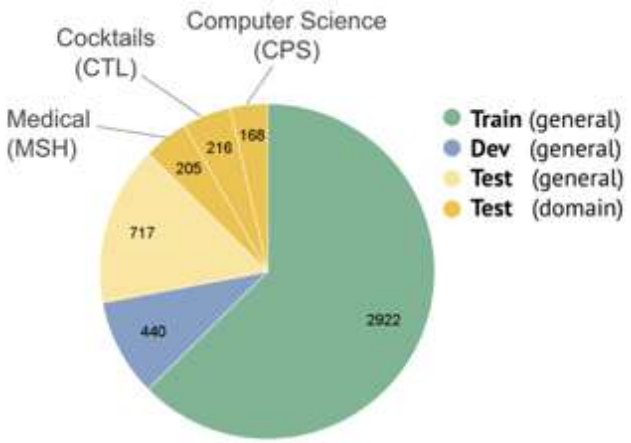

Figure 1. Distribution of training, development, and test set in the WiC-TSV benchmark

Another model quality that the benchmark aims for is the ability to transfer intrinsic knowledge into a specific domain. As for most areas, domain-specific training data is hard to obtain, whereas being able to learn on general-purpose data and still perform well on domain-specific data is a huge advantage in a real-world setting. To simulate this setting, the training and development set of WiC-TSV consist of solely general domain instances adapted from the WIC dataset (Pilehvar and Camacho-Collado 2019), which is based on example phrases from WordNet and Wiktionary, while the test set additionally contains instances from three specific domains, being the computer science, cocktails, and medical domains.

The benchmark further provides a human performance upper bound, which was estimated based on a manually annotated sub-sample of the test set, where the average of the two annotators resulted in an accuracy of 85.5%. This evaluation showed that general purpose

instances seem to be more difficult to solve for humans than the domain-specific ones, as annotators achieved an average accuracy of 82.1% on the general-purpose instances, while the mean accuracy on the domains was 89.1%, 92.0%, and 86.5% for the medical, cocktails, and computer science domains, respectively.

The performance evaluation of a BERT-based (Devlin et al. 2019) baseline model on WiC-TSV shows that current state-of-the-art disambiguation techniques based on pre-trained language models are quite accurate at handling ambiguity, even in specialised domains: the model can achieve an accuracy of 76.3% on the entire test set, and performance on the domain-specific subsets ranged from 72.8% to

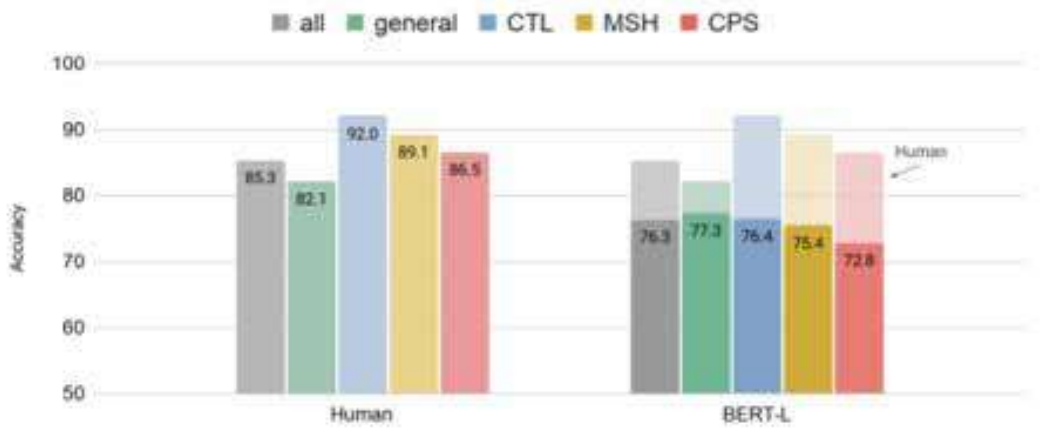

Figure 2. Performance of a BERT-L model on the WiC-TSV benchmark, compared to human performance. Accuracy is provided for the entire test set (all), split into the performance on only the general domain instances (general) and on the performance on the domain-specific subsets (cocktails CTL, medicine MSH, and computer science CPS).

76.4%. However, there is still room for improvement as highlighted by the gap with human performance. This difference in performance also reinforces the challenging nature of the benchmark.

Another point to highlight in this context is the high recall of 82.1 in contrast to its precision of 77.2, which is a desirable characteristic in a retrieval setting as it leads to a high coverage of relevant entities.

Conclusion and future work

The traditional disambiguation task formulations are not suitable for many domain-specific and enterprise settings, as their notion of finding the correct sense in a given sense inventory or knowledge graph restricts the flexibility of disambiguation systems and assumes the availability of all senses. Target Sense Verification is a binary reformulation of the disambiguation task, where the goal is to verify the word in context against one target sense. The advantage of this formulation is its independence of sense inventories and that it does

not assume the availability of all senses, which makes it more suitable for modern domain-specific and enterprise settings. First experiments have shown the capabilities of current state-of-the-art disambiguation models on the presented task, but also highlighted the existing gap that remains until the level of human performance is reached.

For future work, we aim at expanding the existing benchmark, to provide the models with more training examples and be able to evaluate with more comprehensive domain-specific subsets. Furthermore, we plan to extend WiC-TSV to multiple languages, to enable the development of non-English and multilingual disambiguation systems.

References

Auer, A., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Z. Ives. 2007. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC'07/ASWC'07)*, 722-735. Berlin and Heidelberg: Springer-Verlag.

Breit, A., Revenko, A., Rezaee, K., Pilehvar, M.T., and J. Camacho-Collados. 2021. WiC-TSV: An evaluation benchmark for Target Sense Verification of words in context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1635–1645. Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.eacl-main.140/

Camacho-Collados, J. and M.T. Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63, 1: 743–788. https://doi.org/10.1613/jair.1.11259

Devlin, J., Chang, M.-W., Lee, K., and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1: 4171-4186. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N19-1423

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (NIPS'13), 2: 3111–3119. Red Hook, NY: Curran Associates Inc.

Pilehvar, M.T. and J. Camacho-Collado. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1: 1267-1273. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N19-1128

Acknowledgement

This work was partially funded by Prêt-à-LLOD, an EU Horizon 2020 project funded under grant agreement No. 825182. https://pret-a-llod.github.io/