Word Sense-Aware Machine Translation: Including Senses as Contextual Features for Improved Translation Models

Steven Neale[1], Luís Gomes[1], Eneko Agirre[2], Oier Lopez de Lacalle[2] and António Branco[1]
[1]Department of Informatics, Faculty of Sciences
University of Lisbon, Portugal
{steven.neale, luis.gomes, antonio.branco}@di.fc.ul.pt
[2]IXA NLP Group
University of the Basque Country, Spain
{e.agirre, oier.lopezdelacalle}@ehu.eus

Abstract

Although it is commonly assumed that word sense disambiguation (WSD) should help to improve lexical choice and improve the quality of machine translation systems, how to successfully integrate word senses into such systems remains an unanswered question. Some successful approaches have involved reformulating either WSD or the word senses it produces, but work on using traditional word senses to improve machine translation have met with limited success. In this paper, we build upon previous work that experimented on including word senses as contextual features in maxent-based translation models. Training on a large, open-domain corpus (Europarl), we demonstrate that this aproach yields significant improvements in machine translation from English to Portuguese.

Keywords: Machine translation, word senses, maxent-models

1. Introduction

In natural language processing (NLP), word sense disambiguation (WSD) refers to the process of determining the 'sense' or meaning of a word when used in a particular context (Agirre and Edmonds, 2006) – thus solving the common problem of lexical ambiguity in language, where different occurrences of a word token may have multiple distinct meanings. To use a classic example, the word 'bank' could be interpreted in the sense of the financial institution or as the slope of land at the side of a river. Some successful approaches to WSD in recent years have been 'knowledge-based', with classes of words stored in lexical ontologies such as WordNet (Fellbaum, 1998) where the collective meanings of open-class words (nouns, verbs, adjectives and adverbs) are grouped together as 'synsets'.

While it has long been assumed that an optimally successful MT system must incorporate some kind of WSD component (Carpuat and Wu, 2005), attempts to integrate WSD components into machine translation systems have met with mixed – and usually limited – success. Early attempts at 'projecting' word senses directly into a machine translation system (Carpuat and Wu, 2005) were followed by various complete reformulations of the disambiguation process (Carpuat and Wu, 2007; Xiong and Zhang, 2014) – some of which yielded small improvements in translation quality – but the question of whether pure word senses from traditional, knowledge-based WSD approaches can be useful for machine translation still remains.

This paper provides further evidence to support our previous work (Neale et al., 2015), which experimented in including the output from WSD as contextual features in maxent-based translation models in search of improved performance for machine translation from English to Portuguese. Training our transfer model using a much larger dataset – approximately 1.9 million English-Portuguese aligned sentences from Europarl – we find that the very small gains reported previously are now statistically significant, confirming our original hypothesis that adding word senses provided by WSD tools as contextual features of a translation model can improve machine translation performance without the need for intermediary conversion or reformulation of either word senses or the algorithms that deliver them.

We first describe some related work (Section 2) before outlining our approach to integrating the output from WSD into an MT pipeline (Section 3). Next, we describe our evaluation (Section 4) and discuss the results obtained (Section 5), before drawing our conclusions (Section 6).

2. Related Work

Carpuat and Wu (2005) were among the first to challenge the common assumption that WSD should improve upon machine translation performance, demonstrating that many of the contextual features that are important in performing WSD were in fact already implicit in the language models that are trained to perform machine translation. Although acknowledging that the rich semantic data on which dedicated WSD algorithms are based *should* lead to better predictions of lexical choice, they showed that by training a machine translation system on complete parallel sentences they could obtain higher BLEU scores than an equivalent system forcing the output of WSD for isolated words into the translation model (Carpuat and Wu, 2005). These outcomes were to pave the way for a number of subsequent attempts to reformulate the WSD process in such a way as to make it useful for machine translation systems.

The first major work on reformulating WSD for machine translation came from Carpuat and Wu (2007), who proposed that multi-word phrases as opposed to single words be considered as 'phrase-sense disambiguation'. Their approach – leveraging the fact that machine translation models are already trained using contextual features from full sentences – was found to yield improved translation quality across a number of evaluation metrics, suggesting that phrase-based (rather than word-based) sense disambiguation could be more useful to the sentence-based translation models used for machine translation. Further work on reformulating WSD into a more phrase-based concept followed, most notably from Chanel et al. (2007) – who devised a system of creating 'senses' by extracting English translations of full phrases in Chinese and using them as proposed translations – and Giménez and Màrquez (2007) – who used lists of possible translations of single source phrases to predict the correct translations of complete phrases in a given target language.

Renewed interest in exploring the possibility that traditional, single word-based WSD could yet improve machine translation has recently emerged, spearheaded by an approach of integrating pure word senses into machine translation by way of the related technique of 'word sense induction' (WSI) (Xiong and Zhang, 2014). They were successful in predicting the senses of target words – as opposed to predicting their translations, as is the case with phrase-based disambiguation approaches – by clustering words together using their neighbouring words as context to induce ad-hoc 'senses'. This work does, however, leave the question of whether word senses disambiguated using the rich semantic ontologies – such as WordNet – on which traditional WSD is based can be successfully integrated into machine translation pipelines very much open.

In this regard, our own previous work demonstrated a potential step in the right direction by suggesting that incorporating the output from running WSD as contextual features in a maxent-based transfer model results in a slight improvement in the quality of machine translation (Neale et al., 2015). In contrast to two other approaches we had experimented with – 'projecting' word senses into source language input sentences prior to translation either by completely replacing source language lemmas with synset identifiers or appending those synset identifiers onto the source language lemmas – we showed that the reported gains were possible without having to reformulate the word senses themselves nor the algorithms used to retrieve them, as most of the successful marriages of WSD and machine translation reported in the literature had resorted to. While we acknowledged that our work was only in a preliminary state – our reported gains were minimal and based on a very controlled evaluation trained on a small, in-domain corpus – we found any improvement at all to be a good outcome, and recognized the need to continue our experimentation on a larger, open-domain corpus in the future.

3. Integrating WSD Output into the Machine Translation Pipeline

Our chosen WSD algorithm is UKB, a collection of tools and algorithms for performing graph-based WSD over a pre-existing knowledge base (Agirre and Soroa, 2009; Agirre et al., 2014). Based on the graph-based WSD method pioneered by a number of researchers (Navigli and Velardi, 2005; Mihalcea, 2005; Sinha and Mihalcea, 2007; Navigli and Lapata, 2007; Agirre and Soroa, 2008), UKB allows for WordNet-style knowledge bases to be represented as weighted graphs, where word senses correspond to nodes and the relationships or dependencies between pairs of senses correspond to the edges between those nodes. By calculating the probability of a 'random walk' over the graph from a target word's node ending on any other node in the graph – with the nodes (senses) 'recommending' each other and being more or less important based on the importance of the other nodes which recommend them – the most appropriate (probable) sense of the target word in a given context can be chosen (Mihalcea, 2005; Agirre and Soroa, 2009).

We choose to use UKB in our work for two reasons:

- UKB includes tools for automatically creating graph-based representations of knowledge bases in WordNet-style formats.

- The algorithm used by UKB for performing WSD over the graph itself has been consistently shown to produce results in line with or above the state-of-the-art (Agirre and Soroa, 2009; Agirre et al., 2014).

For the purpose of our work, we are thus able to perform highly-efficient WSD over an accurate graph-based representation of our chosen knowledge base (WordNet), meaning that any differences in the results of our various methods of using including senses as contextual features of translation models can be attributed to the methods themselves, rather than to the quality of the WSD output.

3.1. The Domain-Adapted UKB Approach

During the evaluation described in this paper, we also experimented with a domain-adapted approach to performing WSD with UKB. Agirre et al. (2009) demonstrated that knowledge-based WSD can be successfully adapted to specific domains by using information from automatically-built thesauri as context for disambiguation, instead of the original occurrence context of the input sentence. For example, upon finding a domain-specific word in an input sentence we can choose not to provide UKB with the surrounding words in the sentence as context with which to disambiguate the word, but to provide a selection of terms from an automatically-built thesaurus of domain-specific words related to the target word as context instead. This approach has already been shown to be capable of outperforming generic supervised WSD systems (Agirre et al., 2009).

For our evaluations, we produced thesauri calculated from an embedded space obtained using a shallow neural network language model (NNLM) (Mikolov et al., 2013), which have become an important tool in NLP and for semantics in particular in recent years. We used a 'Skip-gram' model, in which each current word is used as the input of a log-linear classifier with a continuous projection layer that predicts the previous and subsequent words within a defined context window. Skip-grams have been shown to be one of the most accurate models available in a variety of semantics-based NLP tasks, such as word similarity and semantic-relatedness (Baroni et al., 2014).

In order to extract the domain-specific thesaurus, an automatically-built corpus of 109 million words was built to provide informative context to the Skip-gram model, comprising 209,000 articles and documents about computer science and information technology extracted from Wikipedia, plus KDE and OpenOffice manuals. The Skip-gram model is then able to learn – following Harris' (1954) distributional hypothesis of a word's semantic features being related to its co-occurrence patterns – word representations as dense scalar vectors of 300 dimensions, with each of these dimensions representing a latent semantic feature. From these 300 dimension embedded vectors, the final thesaurus of each target word is comprised of its 50 most similar words according to cosine similarity.

3.2. Machine Translation System - *TectoMT*

The machine translation system used in our work is TectoMT (Popel and Žabokrtský, 2010), a multi-purpose open source NLP framework that allows for new and existing tools to be created or 'wrapped' as individual modules (known as 'blocks') – reusable and, where possible, language independent – and integrated with each other in various contexts and combinations (known as 'scenarios'). For machine translation, TectoMT breaks down source language input and reconstructs target language output according to four layers of representation: the word layer (raw text), the morphological layer, the analytical layer (shallow-syntax) and the tectogrammatical layer (deep-syntax). The three scenarios needed for machine translation – one for analysis (of the source language), one for transfer (of tectogrammatical nodes from source to target language) and one for synthesis (of the target language) – are constructed from a combination of blocks, for which a pipeline for Portuguese has already been created (Silva et al., 2015).

To integrate word senses into the machine translation pipeline as this paper describes, new combinations of blocks were introduced in the analysis scenario to first convert input sentences into the context format needed to perform WSD using UKB – which we run over a graph-based representation of the English WordNet (Fellbaum, 1998). Another block runs UKB over these sentence-level contexts, returning the appropriate 8-digit synset identifiers for each target word and mapping them back onto the respective word in the source language input. This process happens at the analytical layer, where source language input sentences are represented as a syntactic dependency tree with separate nodes for each word in the sentence – the synset identifiers returned by UKB are added to the nodes of their respective target words as stand-alone attributes.

The final step of integrating word senses is to make use of them as maxent model features, a step which is handled during the transfer scenario of the overall pipeline. After nodes have been encoded from the analytical to the tectogrammatical layer as part of the analysis scenario, the stand-alone synset identifier attributes from different combinations of nodes can be selected for inclusion when computing the maxent model – single target word nodes by themselves, or accompanied by the synset identifiers from their parent nodes (above the target word node in the dependency tree), sibling nodes (to the left and the right of the target node in the dependency tree), or their parent *and* sibling nodes. By including these different combinations of synset identifiers as features in the maxent model used in the transfer of tectogrammatical nodes from source to target language, we expect that the weight added by the synset identifiers would help the maxent model to make better decisions on how individual nodes should be translated.

4. Evaluation

This section describes our currently evaluated methods of including word senses as contextual features in maxent-based translation models. We experiment with three different types of word sense information:

- Traditional word senses (represented by WordNet-stlye synset identifiers).
- 'Supersenses' (represented by the 45 syntactic categories and logical groupings by which synset identifiers are organized in WordNet).
- Both - synset identifiers *and* their corresponding supersense.

We experiment with adding each of these three types of word sense information as features to four different types of node (word) in the maxent-based translation model:

- Node (word).
- Node plus its parent node.
- Node plus its sibling (to the left and right) nodes.
- Node plus its parent *and* sibling nodes.

Finally, we run these experiments using three distinct ways of training our translation models[1]:

- Over a small, in-domain (IT and technological terms) corpus using the regular method of running UKB for WSD.
- Over a small, in-domain corpus using the domain-adapted method of running UKB for WSD.
- Over 1.9 million sentences from the English-Portuguese aligned Europarl corpus using the regular method of running UKB for WSD.

For all evaluations, we analyse the different translation models using a test corpus of 1000 full answers to questions asked by people seeking assistance in resolving problems using technology, as per the domain of the small, in-domain training corpus (section 3.1).

[1] Note that we only experiment with the domain-adapted method of running UKB when training over the small, in-domain corpus, as this method is useful for specific domains and Europarl is here considered to be open-domain.

Baseline: 21.67	**Node**	**+ Parent**	**+ Siblings**	**+ All**
Synset IDs	**21.69**	21.61	**21.68**	21.62
Super-senses	21.64	21.60	21.62	21.58
Both	21.61	21.61	21.63	21.53

Table 1: **Small Corpus and Regular UKB:** A comparison of incorporating word senses – trained over a *small, in-domain corpus* using the *regular approach to running UKB* – as features of different node types in a maxent-based translation model.

Baseline: 21.67	**Node**	**+ Parent**	**+ Siblings**	**+ All**
Synset IDs	**21.68**	21.63	**21.71**	21.65
Super-senses	**21.68**	21.64	21.60	21.64
Both	**21.67**	21.58	21.62	21.54

Table 2: **Small Corpus and Domain-Adapted UKB:** A comparison of incorporating word senses – trained over a *small, in-domain corpus* using the *domain-adapted approach to running UKB* – as features of different node types in a maxent-based translation model.

4.1. Training Over the Small, In-Domain Corpus

For the first two experiments, transfer models were trained over a small, in-domain corpus primarily consisting of 2000 sentences from the QTLeap corpus, a collection of questions and corresponding answers sourced from a real-world company who employ human technicians to provide technical assistance to their customers (technology users) through a chat interface. The QTLeap corpus is described in more detail – in the context of semantic annotation – by Otegi et al. (2016). In addition to these 2000 sentences (1000 questions and 1000 answers), the corpus also includes a number of aligned terms sourced from the localized terminology data of Microsoft (13,000 terms) and LibreOffice (995 terms), making the total size of our in-domain corpus approximately 16,000 paired segments (of which 2000 are full sentences and approximately 14,000 are paired terms).

Table 1 shows our initial experimentation with adding different types of word sense information (either synset IDs, supersenses, or both) retrieved using the regular method of running UKB as contextual features of a) single nodes, b) single nodes plus parent nodes, c) single nodes plus sibling (to the left and right) nodes or 4) singles nodes plus parent *and* sibling nodes used to train maxent-based translation models. The first line of the table, concerning synset IDs, was already reported in our previous work (Neale et al., 2015). The table demonstrates that very small improvements over the baseline BLEU score of 21.67 can be seen when including synset IDs as features of single nodes (21.69) and, to a lesser extent, single nodes plus sibling nodes (21.68). The inclusion of supersenses as features (either by themselves or as well as the synset IDs) bring results below the initial baseline, in most cases significantly so at the 0.05 level.

Table 2 shows the results of the same experimentation but this time using word senses retrieved using the domain-adapted approach to running UKB. The table demonstrates that the same very small improvements over the baseline BLEU score of 21.67 can be seen when including synset IDs as features of single nodes (21.68) and single nodes plus sibling nodes (21.71). Supersenses no longer have a detrimental effect when they are included (either by themselves or as well as the synset IDs) as contextual features of single nodes, yielding very small improvements and suggesting that for single nodes, at least, including any kind of word sense information as contextual features in a maxent-based translation model can improve machine translation output.

4.2. Training Over Europarl

For the third variation of the experiment, transfer models were trained over the Europarl corpus, containing approximately 1.9 million English-Portuguese aligned sentences extracted from the proceedings of the European parliament. Table 3 shows the results of the experiment using this variant, training the same combinations of word sense information and node types as described previously over the full Europarl corpus and using the regular approach to running UKB for performing the WSD. The table demonstrates that by leveraging this considerably larger source of training data, the resulting improvements for machine translation over the baseline BLEU score of 18.31 are now far more widespread and statistically significant at the 0.05 level. – single nodes (18.43) plus parents (18.45) and plus siblings (18.46) for synset IDs, single nodes (18.44) plus siblings (18.44) plus parents *and* siblings (18.46) for supersenses and single plus parent nodes (18.50) and plus siblings (18.41) for both.

5. Discussion

In our previous work (Neale et al., 2015), in which we presented very small (but not significant) improvements over a baseline MT system when adding synset identifiers to single nodes (and to a lesser extent to single nodes plus their sibling nodes), we found any improvement at all to be

Baseline: 18.31	**Node**	**+ Parent**	**+ Siblings**	**+ All**
Synset IDs	**18.43**	**18.45**	**18.46**	18.35
Super-senses	**18.44**	18.30	**18.44**	**18.46**
Both	18.34	**18.50**	**18.41**	18.37

Table 3: **Europarl Corpus and Regular UKB:** A comparison of incorporating word senses – trained over *Europarl* using the *regular approach to running UKB* – as features of different node types in a maxent-based translation model.

in contrast to other approaches we had experimented with, which had left us with results significantly below the baseline. We also acknowledged that our results were based on a very controlled evaluation, trained on a small, in-domain corpus, and that training on larger, open domain corpora such as Europarl might produce different results. Building on this work, the outcomes of this paper can be summarized as follows:

- For training over smaller, in-domain corpora, using the domain-adapted approach to running UKB results in slightly better and slightly more 'compact' results – more results above the baseline and less variation overall, with fewer results significantly below it.
- Training over the larger, open-domain Europarl yields more results that are significantly above the baseline.

While the improvements we report in this paper using the domain-adapted approach to running UKB were still small, there were more results above the baseline and fewer results significantly below the baseline when performing the WSD using the domain-adapted UKB approach. This suggests that the additional, domain-specific context provided to the WSD algorithm using this technique at least helps limit the variance in results between the different methods of including the output of WSD as contextual features. The fact that including supersenses as features – either by themselves or as well as synset identifiers – now results in BLEU scores slightly above the baseline corroborates this theory: we could assume that the extra, domain-specific context would help the correct semantic category (and thus the correct sense) to be selected.

We also acknowledged previously that training on larger, open domain corpora such as Europarl might produce different results, and our second key outcome from the work presented in this paper demonstrates this – training our translation models (using all of our different methods of including the output of WSD as contextual features) over 1.9 million English-Portuguese aligned sentences from Europarl, far more of our results are now over the baseline and significantly so at the 0.05 level of statistical significance. We can also note the emergence of two patterns: an optimal row (synset IDs as the type of word sense information added) and an optimal column (adding word sense information for single nodes plus sibling (to the left and right) nodes) in table 3. We believe that as well as demonstrating the benefits of training transfer models over a larger, open-domain corpus – resulting in a greater coverage of words and therefore, in this case, word senses – these results represent the clearest indication to date that MT output can be improved by including output from WSD directly as contextual features in a maxent-based translation model.

A comparison between the output of the baseline (BLEU score of 18.31) and the highest scoring transfer model (both synset and supersense IDs from the current node and its parent added as features to the maxent model, BLEU score of 18.50) trained over Europarl highlights a number of examples where lexical choice is improved in the model containing word senses. Given a phrase such as "click the right mouse button" (which should translate to "clique com o botão direito do rato"), the baseline model outputs "clique no correcto botão de rato" while the model with word senses outputs "clique no direito botão de rato" – the baseline model has translated the word *right* as *correto* (to be correct) while the model with word senses has made the better lexical choice of *direito* (the opposite of left). Another example is the phrase "allows storage and file creation" (translating to "permite o armazenamento e criação de ficheiros"), for which the baseline model outputs "permite armazenamento e criação de processo" while the model with word senses outputs "permite armazenamento e criação de ficheiro" – the baseline model has translated the word *file* as *processo* (file in the sense of a process) while the model with word senses has made the better lexical choice of *ficheiro* (the Portuguese word typically associated with computer files).

Of course, even the better performing models in our evaluation demonstrate examples of less optimal changes in lexical choice, highlighting the need to continue seeking improvement in the integration of WSD into machine translation. For example, the phrase "you will need to go to the menu Insert > Picture" (translating to "terá de ir ao menu Inserir > Imagem") has been translated by the baseline model as "terá de ir à menu inserção > imagem" and by the model with word senses as "terá de deslocar à menu inserir > imagem". While the model with word senses has delivered one improved lexical choice (*inserir* as opposed to *inserção*) it has also made a less optimal choice in selecting *deslocar* instead of *ir*. This example serves to highlight the delicate interplay between the different types of word sense information and the node types to which they are added, and their subsequent effects on lexical choice.

6. Conclusions

We have presented results that corroborate a hypothesis put forward in previous work that machine translation can be improved by incorporating word senses as contextual features in a maxent-based translation model. Training these models over a large, open domain corpus, we have obtained small but statistically significant improvements in BLEU score (translating from English to Portuguese) when compared to a baseline version of our machine translation system. This demonstrates that including word sense information as features can increase the likelihood of pairings between words and phrases occurring in the translation model.

Our contribution, in showing a statistically significant improvement, is in contrast to previous approaches we have experimented with – namely 'projecting' word senses into source language input sentences prior to translation – that we had found to produce results significantly below our previous baseline (Neale et al., 2015). Furthermore, we interpret our results as evidence that such improvements can be achieved simply using the output of WSD tools, and without the need for any kind of intermediary reformulation or conversion of either the WSD tool itself or its output. This is an important difference between our contribution and previous reports on improving machine translation using word senses, which have tended to involve reformulating the WSD process in some way.

While the improvements we report are statistically significant, they are still small, and we plan further work to see how we could increase the positive effect of including word sense information as contextual features of maxent-based translation models. Notably, we plan to experiment with the mapping word senses to the nodes of source language words using the output of a 'supersense tagger' run over Europarl, as an alternative method of performing the the WSD component. We also acknowledge that it would be worthwhile to explore whether different (perhaps more semantically-oriented) evaluation metrics other than BLEU might provide new insights into our reported improvements using the 'word senses as contextual features' approach.

Acknowledgements

This work has been undertaken and funded as part of the EU project QTLeap (EC/FP7/610516) and the Portuguese project DP4LT (PTDC/EEI-SII/1940/2012).

7. References

Agirre, E. and Edmonds, P. (2006). *Word Sense Disambiguation: Algorithms and Applications*. Springer.

Agirre, E. and Soroa, A. (2008). Using the Multilingual Central Repository for Graph-Based Word Sense Disambiguation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakesh, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41, Athens, Greece. Association for Computational Linguistics.

Agirre, E., Lopez de Lacalle, O., and Soroa, A. (2009). Knowledge-Based WSD on Specific Domains: Performing better than Generic Supervised WSD. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'09)*, pages 1501–1506, Pasadena, USA.

Agirre, E., de Lacalle, O. L., and Soroa, A. (2014). Random Walks for Knowledge-based Word Sense Disambiguation. *Computational Linguistics*, 40(1):57–84, March.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL'14, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Carpuat, M. and Wu, D. (2005). Word Sense Disambiguation vs. Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL-05)*, pages 387–394, Ann Arbor MI, USA.

Carpuat, M. and Wu, D. (2007). How Phrase Sense Disambiguation Outperforms Word Sense Disambiguation for Statistical Machine Translation. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, Skövde, Sweden.

Chan, Y. S., Ng, H. T., and Chiang, D. (2007). Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 33–40, Prague, Czech Republic.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Giménez, J. and Màrquez, L. (2007). Context-Aware Discriminative Phrase Selection for Statistical Machine Translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 159–166, Prague, Czech Republic.

Harris, Z. S. (1954). Distributional Structure. *Word*, 10:146–162.

Mihalcea, R. (2005). Unsupervised Large-vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 411–418, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*.

Navigli, R. and Lapata, M. (2007). Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1683–1688, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Navigli, R. and Velardi, P. (2005). Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1075–1086, July.

Neale, S., Gomes, L., and Branco, A. (2015). First Steps in Using Word Senses as Contextual Features in Maxent Models for Machine Translation. In *Proceedings of the First Workshop on Deep Machine Translation*, DMTW-2015, pages 411–418, Prague, Czech Republic.

Otegi, A., Aranberri, N., Branco, A., Hajič, J., Neale, S., Osenova, P., Pereira, R., Popel, M., Silva, J., Simov, K., and Agirre, E. (2016). QTLeap WSD/NED Corpora: Semantic Annotation of Parallel Corpora in Six Languages. In *Proceedings of the 10th Language Resources and Evaluation Conference*, LREC 2016, Portorož, Slovenia.

Popel, M. and Žabokrtský, Z. (2010). TectoMT: Modular NLP Framework. In *Proceedings of the 7th International Conference on NLP*, IceTal '10, pages 293–304, Reykjavik, Iceland. Springer Berlin Heidelberg.

Silva, J., Rodrigues, J., Gomes, L., and Branco, A. (2015). Bootstrapping a hybrid deep MT system. In *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 1–5, Beijing, China, July. Association for Computational Linguistics.

Sinha, R. and Mihalcea, R. (2007). Unsupervised Graph-basedWord Sense Disambiguation Using Measures of Word Semantic Similarity. In *Proceedings of the International Conference on Semantic Computing*, ICSC '07, pages 363–369, Washington, DC, USA. IEEE Computer Society.

Xiong, D. and Zhang, M. (2014). A Sense-Based Translation Model for Statistical Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL-2014)*, pages 1459–1469, Baltimore MD, USA.