Were the clocks striking or surprising? Using WSD to improve MT performance

Špela Vintar
University of Ljubljana
Dept. of Translation Studies
SI - 1000 Ljubljana,
Aškerčeva 2
spela.vintar@ff.uni-lj.si

Darja Fišer
University of Ljubljana
Dept. of Translation Studies
SI - 1000 Ljubljana,
Aškerčeva 2
darja.fiser@ff.uni-lj.si

Aljoša Vrščaj
University of Ljubljana
Dept. of Translation Studies
SI - 1000 Ljubljana,
Aškerčeva 2
aljosav@gmail.com

Abstract

We report on a series of experiments aimed at improving the machine translation of ambiguous lexical items by using wordnet-based unsupervised Word Sense Disambiguation (WSD) and comparing its results to three MT systems. Our experiments are performed for the English-Slovene language pair using UKB, a freely available graph-based word sense disambiguation system. Results are evaluated in three ways: a manual evaluation of WSD performance from MT perspective, an analysis of agreement between the WSD-proposed equivalent and those suggested by the three systems, and finally by computing BLEU, NIST and METEOR scores for all translation versions. Our results show that WSD performs with a MT-relevant precision of 71% and that 21% of sense-related MT errors could be prevented by using unsupervised WSD.

1 Introduction

Ambiguity continues to be a tough nut to crack in MT. In most known languages certain lexical items can refer to more than a single concept, meaning that MT systems need to choose between several translation equivalents representing different senses of the source word. Wrong choices often result in grave translation errors, as words often refer to several completely unrelated concepts. The adjective *striking* can mean *beautiful, surprising*; *delivering a hard blow* or *indicating a certain time*, and the noun "course" can be *something we give, take, teach* or *eat*.

Our aim was to assess the performance of three MT systems for the English-Slovene language pair and to see whether wordnet-based Word Sense Disambiguation (WSD) could improve performance and assist in avoiding grave sense-related translation errors.

For WSD we use UKB (Agirre and Soroa 2009), a graph-based algorithm that uses wordnet (Fellbaum 1998) and computes the probability of each sense of a polysemous word by taking into account the senses of context words. In our experiment we use Orwell's notorious novel *1984* as the source and its translation into Slovene by Alenka Puhar as the reference translation. We then disambiguate the English source with UKB, assign each disambiguated English word a Slovene equivalent from sloWNet (Fišer 2009) and compare these with the equivalents proposed by Google, Bing and Presis. Results are evaluated in several ways:

- By manually evaluating WSD performance from the MT perspective,
- By analysing the agreement between each of the MT systems and the UKB/wordnet-derived translation,
- By comparing BLEU, NIST and METEOR scores achieved with each translation version.

Our results show that the ad hoc WSD strategies used by the evaluated MT systems can definitely be improved by a proper WSD algorithm, but also that wordnet is not the ideal semantic resource to help resolve translation dilemmas, mainly due to its fine sense granularity.

2 Word Sense Disambiguation and Machine Translation

Wordnet-based approaches to improving MT have been successfully employed by numerous authors, on the one hand as a semantic resource to help resolve ambiguity, and on the other hand as a rich source of domain-specific translation equivalents. As early as 1993 (Knight 1993), wordnet was used as the lower ontology within

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 87–92,
Avignon, France, April 23 - 27 2012.

the PANGLOSS MT system. Yuseop et al. (2002) have employed LSA and the semantic similarity of wordnet literals to translate collocations, while Salam et al. (2009) used wordnet for disambiguation and the choice of the correct translation equivalent in an English to Bengali SMT system.

WSD for machine translation purposes slightly differs from traditional WSD, because distinct source language senses, which share the same translation equivalent, need not be differentiated in WSD (Vickrey et al. 2005). This phenomenon is known as parallel ambiguities and is particularly common among related languages (Resnik and Yarowsky 2000). Although early experiments failed to provide convincing proof that WSD can improve SMT, Carpuat and Wu (2007), Chan et al. (2007) and Ali et al. (2009) clearly demonstrate that incorporating a word sense disambiguation system on the lexical level brings significant improvement according to all common MT evaluation metrics.

Still, using wordnet as the source of sense inventories has been heavily criticized not just in the context of MT (Apidianaki 2009), but also within other language processing tasks. The most notorious arguments against wordnet are its high granularity and - as a consequence - high similarity between some senses, but its global availability and universality seem to be advantages that prevail in many cases (Edmonds and Kilgarriff 2002).

Our experiments lie somewhat in between; on the one hand we demonstrate the potential of WSD in MT, especially for cases where different MT systems disagree, and on the other hand we attribute most WSD errors to the inadequacy of the sense splitting in wordnet (see Discussion).

3 Experimental setup

3.1 Corpus and MT systems

Our corpus consists of George Orwell's novel *1984*, first published in English in 1949, and its translation into Slovene by Alenka Puhar, first published in 1967. While it may seem unusual to be using a work of fiction for the assessment of MT systems, literary language is usually richer in ambiguity and thus provides a more complex semantic space than non-fiction.

We translated the entire novel into Slovene with Google Translate[1], Bing[2] and Presis[3], the first two belonging to the family of freely available statistical systems and the latter being a rule-based MT system developed by the Slovenian company Amebis.

For the purposes of further analysis and comparison with our disambiguated corpus all texts - original and translations - have been PoS-tagged and lemmatized using the JOS web service (Erjavec et al. 2010) for Slovene and ToTaLe (Erjavec et al. 2005) for English. Because we can only disambiguate content words, we retained only nouns, verbs, adjectives and adverbs and discarded the rest. After all these preprocessing steps our texts end up looking as follows:

***English**:*
It was a bright cold day in April and the clocks were striking thirteen.
***English-preprocessed**:*
be bright cold day April clock be strike
Slovene-reference:
Bil je jasen, mrzel aprilski dan in ure so bile trinajst.
Slovene-reference-preprocessed:
biti biti jasen mrzel aprilski dan ura biti biti
Slovene-Google:
Bilo je svetlo mrzel dan v aprilu, in ure so bile trinajst presenetljiv.
Slovene-Google-preprocessed:
biti biti svetlo mrzel dan april ura biti biti presenetljiv
Slovene-Bing:
Je bil svetlo hladne dan aprila in v ure so bili presenetljivo trinajst.
Slovene-Bing-preprocessed:
biti biti svetlo hladen dan april ura biti biti presenetljivo
Slovene-Presis:
Svetal hladen dan v aprilu je bilin so ure udarjale trinajst.
Slovene-Presis-preprocessed:
svetel hladen dan april biti bilin biti ura udarjati

Figure 1. Corpus preprocessing

3.2 Disambiguation with UKB and wordnet

The aim of semantic annotation and disambiguation is to identify polysemous lexical items in the English text and assign them the correct sense in accordance with the context. Once the sense of the word has been determined, we can exploit the cross-language links between wordnets of different languages and propose a Slovene translation equivalent from the Slovene wordnet.

We disambiguated the English corpus with UKB, which utilizes the relations between synsets and constructs semantic graphs for each candidate sense of the word. The algorithm then

[1] http://translate.google.com (translation from and into Slovene has been available as of September 2008)

[2] http://www.microsofttranslator.com/ (available for Slovene since 2010)

[3] http://presis.amebis.si (available for English-Slovene since 2002)

computes the probability of each graph based on the number and weight of edges between the nodes representing semantic concepts. Disambiguation is performed in a monolingual context for single- and multiword nouns, verbs, adjectives and adverbs, provided they are included in the English wordnet.

Figure 2 shows the result of the disambiguation algorithm for the word *face,* which has as many as 13 possible senses in wordnet. We are given the probability of each sense in the given context (eg. *0.173463*) and the ID of the synset (eg. *eng-30-05600637-n*), and for the purposes of clarity we also added the literals (words) associated with this particular synset ID in the English (*face, human face*) and Slovene (*fris, obraz, faca*) wordnet respectively. As can be seen from this example, wordnet is - in most cases - a very fine-grained sense inventory, and looking at the Slovene equivalents clearly shows that many of these senses may partly or entirely overlap, at least in the context of translation.

WSD: ctx_Oen.1.1.2 24 !! face

- *W: 0.173463 ID: eng-30-05600637-n ENGWN: face, human face, (the front of the human head from the forehead to the chin and ear to ear) SLOWN: fris, obraz, faca, človeški obraz, (EMPTYDEF)*
- *W: 0.116604 ID: eng-30-08510666-n ENGWN: side, face, (a surface forming part of the outside of an object) SLOWN: stranica, ploskev, (EMPTYDEF)*
- *W: 0.0956895 ID: eng-30-03313602-n ENGWN: face, (the side upon which the use of a thing depends (usually the most prominent surface of an object)) SLOWN: sprednja stran, prava stran, zgornja stran, lice, (EMPTYDEF)*
- *W: 0.0761554 ID: eng-30-04679738-n ENGWN: expression, look, aspect, facial expression, face, (the feelings expressed on a person's face) SLOWN: izraz, pogled, obraz, izraz na obrazu, (EMPTYDEF)*
- *W: 0.0709513 ID: eng-30-03313456-n ENGWN: face, (a vertical surface of a building or cliff) SLOWN: stena, fasada, (EMPTYDEF)*
- *W: 0.0653514 ID: eng-30-06825399-n ENGWN: font, fount, typeface, face, case, (a specific size and style of type within a type family) SLOWN: font, pisava, črkovna družina, vrsta črk, črkovna podoba, črkovni slog, (EMPTYDEF)*
- *W: 0.0629878 ID: eng-30-04838210-n ENGWN: boldness, nerve, brass, face, cheek, (impudent aggressiveness) SLOWN: predrznost, nesramnost, (EMPTYDEF)*
- *W: 0.0610286 ID: eng-30-06877578-n ENGWN: grimace, face, (a contorted facial expression) SLOWN: spaka, grimasa, (EMPTYDEF)*
- *W: 0.0605221 ID: eng-30-03313873-n ENGWN: face, (the striking or working surface of an implement) SLOWN: čelo, podplat, udarna površina, (EMPTYDEF)*
- *W: 0.0579952 ID: eng-30-05601198-n ENGWN: face, (the part of an animal corresponding to the human face) SLOWN: obraz, (EMPTYDEF)*
- *W: 0.0535548 ID: eng-30-05168795-n ENGWN: face, (status in the eyes of others) SLOWN: ugled, dobro ime, (EMPTYDEF)*
- *W: 0.05303 ID: eng-30-09618957-n ENGWN: face, (a part of a person that is used to refer to a person) SLOWN: obraz, (EMPTYDEF)*
- *W: 0.0526668 ID: eng-30-04679419-n ENGWN: face, (the general outward appearance of something) SLOWN: podoba, (EMPTYDEF)*

Figure 2. Disambiguation result for the word *face* with probabilities for each of the twelve senses

As can be seen in Table 1, almost half of all the tokens in the corpus are considered to be ambiguous according to the English wordnet. Since the Slovene wordnet is considerably smaller than the English one, almost half of the different ambiguous words occurring in our corpus have no equivalent in sloWNet. This could affect the results of our experiment, because we cannot evaluate the potential benefit of WSD if we cannot compare the translation equivalent from sloWNet with the solutions proposed by different MT systems. We therefore restricted ourselves to the words and sentences for which an equivalent exists in sloWNet.

Corpus size in tokens	103,769
Corpus size in types	10,982
Ambiguous tokens	48,632
Ambiguous types	7,627
Synsets with no equivalent in sloWNet	3,192

Table 1. Corpus size and number of ambiguous words

One method of evaluating the performance of WSD in the context of Machine Translation is through metrics for automatic evaluation (BLEU, NIST, METEOR etc.). We thus generated our own translation version, in fact a stripped version similar to those in Figure 1 consisting only of content words in their lemmatized form. We translated the disambiguated words with wordnet, exploiting the cross-language universality of the synset ID. However, since we can only propose translation equivalents for the words which are included in wordnet, we had to come up with a translation solution for those which were not. Such words include proper names (*Winston, Smith, London, Oceania*), hyphenated compounds (*pig-iron, lift-shaft, gorilla-faced*) and Orwellian neologisms (*Minipax, Newspeak, thoughtcrime*). We translated these words with three alternative methods:

- Using a general bilingual dictionary,
- Using the English-Slovene Wikipedia and Wiktionary,

- Using the automatically constructed bilingual lexicon from the English-Slovene parallel Orwell corpus.

The fourth option was to leave them untranslated and simply add them to the generated Slovene version.

4 Evaluation

The number of meanings a word can have, the degree of translation equivalence or the quality of the target text are all extremely disputable and vague notions. For this reason we wished to evaluate our results from as many angles as possible, both manually and automatically.

4.1 Manual evaluation of WSD precision in the context of MT

Firstly, we were interested in the performance of the UKB disambiguation tool in the context of MT. Since UKB uses wordnet as a sense inventory, the algorithm assigns a probability to each sense of a lexical item according to its context in an unsupervised way. The precision of UKB for unsupervised WSD is reported at around 58% for all words and around 72% for nouns, but of course these figures measure the number of cases where the algorithm selected the correct wordnet synset from a relatively fine-grained network of possible senses (Agirre and Soroa 2009).
We adjusted the evaluation task to an MT scenario by manually checking 200 disambiguated words and their suggested translation equivalents, and if the equivalent was acceptable we counted it among the positive instances regardless of the selected sense. For example, the English word *breast* has four senses in wordnet: (1) the upper frontal part of a human chest, (2) one of the two soft milk-secreting glands of a woman, (3) meat carved from the breast of a fowl and (4) the upper front part of an animal corresponding to the human chest. For the English sentence *Winston nuzzled his chin into his breast*... UKB suggested the second sense, which is clearly wrong, but since the ambiguity is preserved in Slovene and the word *prsi* can be used for all of the four meanings, we consider this a case of successful disambiguation for the purposes of MT.

Translation equivalent	correct	incorrect	borderline
Number/ %	142 (71%)	46 (23%)	12 (6%)

Table 2: Manual evaluation of WSD performance for MT

The precision of WSD using this relaxed criterion was 71%, with 6% so-called borderline cases. These include cases where the equivalent was semantically correct but had the wrong part of speech (eg. ***glass*** *door* -> ****steklo*** instead of ***steklen***).

4.2 Agreement between each of the MT systems and the disambiguated equivalent

It is interesting to compare the equivalents we propose through our wordnet-based WSD procedure with those suggested by the three MT systems: Presis, Google and Bing.

Total no. of disambiguated tokens	13,737
WSD = reference	3,933
WSD = Presis	4,290
WSD = Google	4,464
WSD = Bing	4,377
WSD = ref = Presis = Google = Bing	2,681
WSD = ref ≠ Presis ≠ Google ≠ Bing	269

Table 3: Comparison of WSD/wordnet-based equivalent and the translations proposed by Presis, Google, Bing and the reference translation

The comparison was strict in the sense that we only took into account the first Slovene equivalent proposed within the same synset. Of the over 48k ambiguous tokens we obviously considered only those which had an equivalent in sloWNet, otherwise comparison with the MT systems would have been impossible. We can see from Table 2 that the WSD/wordnet-based equivalents most often agree with Google translation, and that for approximately every fifth ambiguous word all systems agree with each other and with the reference translation.

If we also look at the number of cases where our WSD-wordnet-based equivalent is the only one to agree with the reference translation, it is safe to assume that these are the cases where WSD could clearly improve MT. Of all the instances where WSD agrees with the reference translation we can subtract the instances where all systems agree, because these need no improvement. Of the remaining 1,252 ambiguous words, 269 or 20% were such that only the WSD/wordnet equivalent corresponded to the reference translation.

4.3 Evaluation with metrics

Finally, we wanted to see how the WSD/wordnet-based translation compares with the three MT systems using the BLEU, NIST and METEOR scores. For the purposes of this comparison we pre-processed all five versions of our corpus - original, reference translation, Presis, Google and Bing translation - by lemmatization, removal of all function words, removal of sentences where the alignment was not 1:1, and finally by removal of the sentences which contained lexical items for which there was no equivalent in sloWNet.

We then generated the sixth version by translating all ambiguous words with sloWNet (see Section 3), and for the words not included in the English wordnet we used four alternative translation strategies; a general bilingual dictionary (dict), wiktionary (wikt), a word-alignment lexicon (align) and amending untranslated words to the target language version (amend).

	BLEU (n=1)	NIST	METEOR
Bing	0.506	3.594	0.455
Google	0.579	4.230	0.481
Presis	0.485	3.333	0.453
WSD	0.440	3.258	0.429
WSD-amend	0.410	3.308	0.430
WSD-dict	0.405	3.250	0.427
WSD-align	0.448	**3.588**	0.434
WSD-wikt	0.442	3.326	0.429

Table 4: Evaluation with metrics

Table 3 shows the results of automatic evaluation; the corpus consisted of 2,428 segments. We can see that our generated version using disambiguated equivalents does not outperform any of the MT systems on any metric, except once when the WSD-align version outperforms Presis on the NIST score and comes fairly close to the Bing score.

It is possible that the improvement we are trying to achieve is difficult to measure with these metrics because our method operates on the level of single words, while the metrics typically evaluate entire sentences and corpora. We are using a stripped version of the corpus, ie. only content words which can potentially be ambiguous, whereas the metrics are normally used to calculate the similarity between two versions of running text. Finally, the corpus we are using for automatic evaluation is very small.

5 Discussion

Although employing WSD and comparing wordnet-based translation equivalents to those proposed by MT systems scored no significant improvement with standard MT evaluation metrics, we remain convinced that the other two evaluation methods show the potential of using WSD, particularly with truly ambiguous words and not those where sense distinctions are slight or vague. A manual inspection of the examples where MT systems disagreed and our WSD-based equivalent was the only one to agree with the reference translation shows that these are indeed examples of grave MT errors. For example, the word *hand* in the sentence *The clock's hands said six meaning eighteen* can only be translated correctly with a proper WSD strategy and was indeed mistranslated as *roka* (body part) by all three systems. If a relatively simplistic and unsupervised technique such as the one we propose can prevent 20% of these mistakes, it is certainly worth employing at least as a post-processing step.

The fact that we explore the impact of WSD on a work of fiction rather than domain-specific texts may also play a role in the results we obtained, although it is not entirely clear in what way. We believe that in general there is more ambiguity in literary texts meaning that a single word will appear in a wider range of senses in a work of fiction than it would in a domain-specific corpus. This might mean that WSD for literary texts is more difficult, however our own experiments so far show no significant difference in WSD performance.

A look at the cases where WSD goes wrong shows that these are typically words with a high number of senses which are difficult to differentiate even for a human. The question from the title of this paper is actually a translation blunder made by both Google and Bing, since *striking* was interpreted in its more expressive sense and translated into Slovene as *presenetljiv [surprising]*. However, UKB also got it wrong and chose the sense defined as *deliver a sharp blow, as with the hand, fist, or weapon* instead of *indicate a certain time by striking.* While these meanings may seem quite easy to tell apart, especially if the preceding word in a sentence is *clock*, *strike* as a verb has as many as 20 senses in Princeton WordNet, and many of these seem very similar. In this case the Slovene translation we propose is "less wrong" than the *surprising* solution offered by Google or Bing, because *udarjati* may actually be used in the *clock* sense as well.

We might also assume that statistical MT systems will perform worse on fiction; results in Table 3 show that both statistical systems outperform the rule-based Presis. Then again, Orwell's 1984 has been freely available as a parallel corpus for a very long time and it is therefore possible that both Google and Bing have used it as training data for their SMT model.

6 Conclusion

We described an experiment in which we explore the potential of WSD to improve the machine translation of ambiguous words for the English-Slovene language pair. We utilized the output of UKB, a graph-based WSD tool using wordnet, to select the appropriate equivalent from sloWNet. Manual evaluation showed that the correct equivalent was proposed in 71% of the cases. We then compared these equivalents with the output of three MT systems. While the benefit of WSD could not be proven with the BLEU, NIST and METEOR scores, the correspondence of the WSD/wordnet-based equivalent with the reference translation was high. Furthermore it appears that in cases where MT systems disagree WSD can help choose the correct equivalent.

As future work we plan to redesign the experiment so as to directly use WSD as a post-processing step to machine translation instead of generating our own stripped translation version. This would provide better comparison grounds. In order to improve WSD precision we intend to combine two different algorithms and use it only in cases where both agree. Also, we intend to experiment with different text types and context lengths to be able to evaluate WSD performance in the context of MT on a larger scale.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. Proceeding of the European Association of Computational Linguistics conference (EACL09).

Ola Mohammad Ali, Mahmoud Gad Alla and Mohammad Said Abdelwahab. 2009. Improving machine translation using hybrid dictionary-graph based word sense disambiguation with semantic and statistical methods. International Journal of Computer and Electrical Engineering, 1/5.

Marianna Apidianaki. 2009. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. Proceedings of the 12th Conference of the European Chapter of the ACL, pages 77–85, Athens, Greece, Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. Proceedings of Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).

Yee Seng Chan, Hwee Tou Ng and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (Prague, Czech Republic). 33-40.

Philip Edmonds and Adam Kilgarriff. 2002. Introduction to the Special Issue on Evaluating Word Sense Disambiguation Systems. Natural Language Engineering 8 (4): 279–291.

Tomaž Erjavec, Darja Fišer, Simon Krek and Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10), Malta.

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

Darja Fišer. 2009. Leveraging Parallel Corpora and Existing Wordnets for Automatic Construction of the Slovene Wordnet. Human language technology: challenges of the information society, (LNCS 5603). Berlin; Heidelberg: Springer: 359-368.

Kevin Knight. 1993. Building a large ontology for machine translation. Proceedings of the ARPA Human Language Technology Workshop, Plainsboro, New Jersey.

Philip Resnik and David Yarowsky. 2000. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. Natural Language Engineering, 5(2): 113-133.

Khan Md. Anwarus Salam, Mumit Khan and Tetsuro Nishino. 2009. Example based English-Bengali machine translation using wordnet. Proceedings of TriSA'09, Japan.

David Vickrey, Luke Biewald, Marc Teyssier in Daphne Koller. 2005. Word-Sense Disambiguation for Machine Translation. Proceedings of the Conference Empirical Methods in Natural Language Processing (EMNLP).

Kim Yuseop, Jeong-Ho Chang in Byoung-Tak Zhang (2002): Target Word Selection Using WordNet and Data-Driven Models in Machine Translation. Proceedings of the Conference PRICAI'02: Trends in Artificial Intelligence.