Data-driven semantic analysis for multilingual WSD and lexical selection in translation

Marianna Apidianaki
National Centre for Language Technology
School of Computing, Dublin City University
Dublin 9, Ireland
mapidianaki@computing.dcu.ie

Abstract

A common way of describing the senses of ambiguous words in multilingual Word Sense Disambiguation (WSD) is by reference to their translation equivalents in another language. The theoretical soundness of the senses induced in this way can, however, be doubted. This type of cross-lingual sense identification has implications for multilingual WSD and MT evaluation as well. In this article, we first present some arguments in favour of a more thorough analysis of the semantic information that may be induced by the equivalents of ambiguous words found in parallel corpora. Then, we present an unsupervised WSD method and a lexical selection method that exploit the results of a data-driven sense induction method. Finally, we show how this automatically acquired information can be exploited for a multilingual WSD and MT evaluation more sensitive to lexical semantics.

1 Word senses in a bi-(multi-)lingual context

1.1 Cross-lingual sense determination for WSD

Determining the senses of ambiguous words by reference to their translational equivalents constitutes a common practice in multilingual WSD: the candidate senses of an ambiguous word, from which one has to be selected during WSD, correspond to its equivalents in another language. This empirical approach to sense identification circumvents the need for predefined sense inventories and their disadvantages for automatic WSD.[1] The first to adopt it were Brown *et al.* (1991), who represented the two main senses of a SL word by its two most frequent translations in the target language (TL). Further promoted by Resnik and Yarowsky (2000) and endorsed in the multilingual tasks of the Senseval (Chklovski *et al.*, 2004) and Semeval (Jin *et al.*, 2007) exercises, this conception of senses is still found in recent works on the integration of WSD in MT.

From these works, only that of Carpuat and Wu (2005) exploits an external hand-crafted sense inventory. The use of an external resource, not related to the training corpus of their Statistical Machine Translation (SMT) system, turned out to be one of the causes of the observed deterioration of translation quality. In later works on the subject, which show a more or less important improvement in translation quality, SL word senses are considered as directly reflected in their equivalents found in a parallel training corpus (Cabezas and Resnik, 2005; Carpuat and Wu, 2007; Chan *et al.*, 2007). Nevertheless, the theoretical soundness of these senses is not really addressed.

1.2 Advantages of cross-lingual sense determination

Cross-lingual sense induction offers a standard criterion for sense delimitation: the translation equivalents of ambiguous words are supposed to reveal their hidden meanings (Resnik, 2004). Additional advantages become evident in MT: when the candidate senses of an ambiguous word consist of its possible translations, identifying the sense carried by a new instance of the word coincides with its translation. Conceiving WSD

[1] Such as the high granularity, the great number and the striking similarity of the described senses, and their irrelevance to the domains of the processed texts (Edmonds and Kilgarriff, 2002).

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 77–85,
Athens, Greece, 30 March – 3 April 2009.

as lexical selection thus seems natural (Vickrey *et al*., 2005): it appears that there is no reason to pass through senses in order to arrive to translations. A correct translation may be attained even without WSD, as in the case of parallel ambiguities where the SL and TL words are similarly ambiguous (Resnik and Yarowsky, 2000).[2]

1.3 Disadvantages of cross-lingual sense determination

However, this conception of senses is not theoretically sound, as translation equivalents do not always constitute valid sense indicators. This is often neglected in an attempt to render the sense inventory as close as possible to the training corpus of the SMT system. So, translation equivalents are considered as straightforward indicators of SL senses.

This approach assumes and results in some type of uniformity regarding the nature of the induced senses: clear-cut (e.g. homonymic) and finer sense distinctions are all handled in the same way. Moreover, senses are enumerated without any description of their possible relations. For instance, a SL word *w* having three equivalents (*a*, *b* and *c*) is considered to have three distinct senses (described as '*w*-*a*', '*w*-*b*' et '*w*-*c*)'.

The assumption of biunivocal (one-to-one) correspondences between senses and equivalents disregards the fact that semantically similar equivalents may be used to translate the same sense of a SL word in context. However, this constitutes a common practice in translation and an advised technique for translators, in order to avoid repetitions in the translated texts. The phenomenon of translation ambiguity may pose some problems as well: it may not need to be resolved during translation but should be considered in multilingual WSD. Resolving this kind of ambiguity could also improve the quality of the results of applications such as multilingual information retrieval.

1.4 Impact of cross-lingually defined senses on evaluation

Ignoring the relations between word senses may raise further problems during WSD evaluation, as errors concerning close or distant senses are considered as equally important. Thus, if a WSD algorithm selects a sense which is slightly different from the one effectively carried by an instance of an ambiguous word, but not totally wrong, this is directly considered as a false choice. A differing weighting of WSD errors would be preferable in these cases, if sense distance information was available (Resnik and Yarowsky, 2000).

When WSD coincides with lexical selection in MT, the equivalents of a SL word (*w*) are perceived to be its candidate senses. The sense assigned to a new instance of *w* is considered to be correct if it corresponds to the reference translation (i.e. the translation of that instance in the test corpus). This strict requirement of exact correspondence constitutes one of the main critics addressed to MT evaluation metrics (Cabezas and Resnik, 2005; Callison-Burch, 2006; Chan *et al*., 2007) and is one of the main reasons that methods have been developed which go beyond pure string matching (Owczarzak *et al*., 2007).

A central issue in MT evaluation is the high correlation of the metrics with human judgements of translation quality, which puts the accent on the identification of sense correspondences. Here too, it is essential to penalize errors relatively to their importance and so information relative to the semantics of the equivalents should be available. In the next section we will show how this information can be acquired using a data-driven sense induction method.

2 Data-driven semantic analysis in a bilingual context

We propose to explore the semantic relations of the equivalents of ambiguous words using a parallel corpus and to exploit these relations for SL sense induction. A data-driven sense acquisition method based on this type of relations is presented in Apidianaki (2008). The theoretical assumptions underlying this approach are the distributional hypotheses of meaning (Harris, 1954) and of semantic similarity (Miller and Charles, 1989), and that of sense correspondence between words in translation relation in real texts.

Our training corpus is the English (EN)–Greek (GR) part of the lemmatized and POS-tagged INTERA corpus (Gavrilidou *et al*., 2004) which contains approximately four million words. The corpus has been sentence- and word-aligned at the level of tokens and types (Simard and Langlais, 2003). Two bilingual lexicons (one

[2] A typical example is that of the ambiguous English noun *interest* whose "personal" and "financial" senses are translated by the same word in French (*intérêt*).

for each translation direction: EN–GR/GR–EN are built from the alignment of word types. In these lexicons, each SL word (*w*) is associated with the set of the equivalents to which it is aligned, as shown hereafter:

implication: {*συνέπεια* (consequence), *επίπτωση* (impact), *επιπλοκή* (complication)}

variation: {*διακύμανση* (fluctuation), *μεταβολή* (alteration), *τροποποίηση* (modification)}

The words in parentheses describe the senses of the Greek equivalents. In order to eliminate the noise present in the lexicons, two filters are used: a POS-filter, that keeps only the correspondences between words of the same category[3] and an intersection filter, which discards the translation correspondences not found in both translation lexicons. A lexical sample of 150 ambiguous English nouns having more than two equivalents is then created from the EN–GR lexicon[4]. At this stage, the semantic relations possibly existing between the equivalents are not yet evident, and so no conclusions can be extracted concerning the distinctiveness of the senses they can induce on the SL words.

The core component of the sense induction method used is a semantic similarity calculation which aims at discovering the relations between the equivalents of a SL ambiguous word (*w*). First, the translation units (TUs)[5] in which *w* appears in the SL sentence(s) are extracted from the training corpus and are then grouped by reference to *w*'s equivalents. For instance, if *w* is translated by *a*, *b* and *c*, three sets of TUs are formed (where *w* is translated by *a* ('*w*-*a*' TUs), by *b* ('*w*-*b*' TUs), etc.).

The SL context features corresponding to each equivalent (i.e. the set of lemmatized content words surrounding *w* in the SL side of the TUs corresponding to the equivalent) are extracted and treated as a 'bag of words'. This distributional information serves to calculate the equivalents' similarity using a variation of the Weighted Jaccard coefficient (Grefenstette, 1994). The similarity calculation is described in detail in Apidianaki (2008).

Each retained context feature is assigned a weight relatively to each equivalent, which serves to define its relevance for the estimation of the equivalents' similarity. The equivalents are compared in a pairwise manner and a similarity score is assigned to each pair. Two equivalents are considered as semantically related if the instances of *w* they translate in the training corpus occur in "similar enough" contexts. The pertinence of their relation is judged by comparing its score to a threshold, equal to the mean of the scores assigned to all the pairs of equivalents of *w*.

The results of this calculation are exploited by a clustering algorithm which takes as input the set of equivalents of *w* and outputs clusters of similar equivalents illustrating its senses (Apidianaki, 2008). Clustered equivalents are semantically related[6] and considered as translating the same SL sense, while isolated ones translate distinct senses.

The same calculation is performed by reference to the TL contexts of the equivalents, i.e. using the lemmatized content words surrounding the equivalents in the TL side of the corresponding TUs sets. Contrary to the SL results, the TL ones are not used for clustering. The TL distributional information relative to the clustered equivalents and acquired at this stage will be used for lexical selection, as we will show later in this paper.

The sense clusters created for a word serve to identify its senses. We describe the senses acquired for the nouns *implication* and *variation*:

implication:
{*συνέπεια*, *επίπτωση*}: the "impact" sense
{*επιπλοκή*}: the "complication" sense

variation:
{*διακύμανση*}: the "fluctuation" sense
{*μεταβολή, τροποποίηση*}: the "alteration" sense

The sense induction method presented above thus permits the automatic creation of a sense inventory from a parallel corpus. In what follows, we will show how this can be exploited for WSD.

3 Unsupervised WSD based on the semantic clustering

The method described in section 2 provides, as a by-product, information that can be exploited by an unsupervised WSD classifier. In the case of a one-equivalent cluster, this information corresponds to the set of the equivalent's

[3] The noun equivalents of nouns, the verb equivalents of verbs, etc.

[4] Here we focus on nouns but the method is applicable to words of other POS categories.

[5] A translation unit contains up to 2 sentences of each language linked by an alignment.

[6] Most often near-synonyms but they may be linked by other relations (hyperonymy, hyponymy, etc.).

features, retained from the corresponding SL contexts of *w*. In the case of bigger clusters, it consists of the SL context features that reveal the equivalents' similarities: for a cluster of two equivalents, it consists of their assimilative contexts (i.e. the features they share)[7]; for a cluster of more than 2 equivalents, it consists of the intersection of the common features of the pairs of equivalents found in the cluster.

As we have already said, each retained context feature is assigned a weight relatively to each equivalent. Here are the weighted features characterizing the clusters of *variation* :

{διακύμανση}: *significant* (2.04), *range* (0.76), p*harmacokinetics*(1.89), *individual* (1.89), *affect* (1.89), *insulin* (1.89), *woman* (1.89), *year* (1.49), *man* (1.19), *considerable* (1.19), *member* (1.12), *old* (0.76), *Ireland* (0.76), *case* (0.72), *increase* (0.76), *group* (0.76), *states* (0.71), *external* (0.76), *good* (0.76), *expectancy* (0.76), *Spain* (0.76), *pressure* (0.76), *Europe* (0.76)

{τροποποίηση, μεταβολή} : *minor* (2.25/1.83), *human* (2.01/1.13), *number* (0.73/1.16) [8]

In order to disambiguate a new instance of a word *w*, cooccurrence information coming from its context is compared to the sets of features characterizing the clusters. The new context must thus be lemmatized and POS-tagged as well. Here is an example of a new instance of *variation*:

a. "*Although certain regions have been faced with an exodus of their endogenous population, most of the coastal zones are experiencing an increase in overall demographic pressure, as well as significant seasonal* ***variations*** *in employment, essentially linked to tourism.*"

The features retained from this context are the lemmas of the content words (nouns, verbs and adjectives) surrounding *w*. If common features (CFs) are found between this context and just one cluster of *w*, this is selected as describing the sense of the new instance. On the contrary, if CFs with more than one cluster are found, a score is given to each *context-cluster* association. This score corresponds to the mean of the weights of the CFs relatively to each equivalent of the cluster and is given by the following formula.

$$\frac{\sum_{i=1}^{e}\sum_{j=1}^{f} w\,(equivalent_i\,,\,feature_j)}{e * f}$$

In this formula, e is the number of the equivalents of a cluster and f is the number of its CFs with the new context. The cluster with the highest score is retained; it describes the sense carried by the new instance *of w* and could be used as its sense tag. The only cluster having CFs with the context of *variation* in (a) and is thus selected is {*διακύμανση*} (CFs: *increase*, *pressure, significant*).

If any instances remain ambiguous at the end of the WSD process (i.e. no associations are established with the sense clusters), a small modification could increase the method's coverage. If *w* has clusters of more than two equivalents, it is possible to use the assimilative contexts of the pairs of equivalents instead of their intersection. The coverage of the WSD method would be increased in this way, as the sets of assimilative contexts would contain more features than their intersection, and so it would become more probable to find CFs with the new contexts and to establish '*context-cluster*' associations.

4 Semantics-sensitive WSD evaluation

4.1 The notion of *enriched precision*

In this section, we will present the evaluation of the proposed WSD method and we will show how the clustering information can be exploited at this stage.[9] The new instances of the nouns of our lexical sample, used for evaluation, come from our test corpus, the sentence aligned EN–GR part of EUROPARL (Koehn, 2005). The TUs containing the ambiguous nouns are extracted from the corpus. Lacking a gold-standard for evaluation, we exploit information relative to translations.

In the multilingual tasks of Senseval and Semeval (Ckhlovski *et al*., 2004; Jin *et al*., 2007), the translations of the words in the parallel test corpus are considered as their sense tags. Here, we consider that the equivalent translating an ambiguous SL word in context (called *reference translation*) points to a sense described by a cluster. Consequently, what is being evaluated is the capacity of the WSD

[7] Term used in the study of paraphrase (Fuchs, 1994).

[8] The two scores in parentheses correspond, respectively, to the score of the feature by reference to the first and the second equivalent of the cluster.

[9] Some of the equivalents of *w* found in the training corpus and contained in the clusters may not be used in the test corpus. The evaluation concerns only those that are found in the test corpus.

method to predict this sense. The sense proposed for an instance of an ambiguous word is considered as **correct** if a) a 1-equivalent cluster is selected and the equivalent corresponds to the reference, or b) if a bigger cluster containing the reference is selected. Otherwise, the proposed sense is **false**.

In the multilingual tasks where translations are regarded as sense tags, the proposed senses are considered as correct only if they correspond exactly to the reference translation. This is the principle of *precision*, underlying most of the existing MT evaluation metrics. From a quantitative point of view, this strict criterion has a negative impact on the WSD evaluation results. From a qualitative point of view, it ignores the fact that different equivalents may correspond to the same source sense and that an ambiguous word in context can have more than one good translation.

The use of the sense clusters during WSD evaluation offers the possibility of capturing the semantic relations between the equivalents of ambiguous words, acquired during learning. In this case, the evaluation could be considered as based on a principle of *enriched precision* that exploits the paradigmatic relations of TL words.

4.2. Evaluation metrics

The metrics used for WSD evaluation are the following:

$$recall = \frac{number\ of\ correct\ predictions}{number\ of\ new\ instances}$$

$$precision = \frac{number\ of\ correct\ predictions}{number\ of\ predictions}$$

The obtained results are compared to those of a baseline method. The baseline most often used in Senseval is that of the most frequent sense (i.e. the first sense given for a word in a predefined sense inventory). This is a very powerful heuristic because of the asymmetric distribution of word senses in real texts. Our baseline consists of choosing the most frequent equivalent (i.e. the one that translates w most frequently in the training corpus) as illustrating the sense of all its new instances. The asymmetric distribution of senses is, however, reflected at the level of the equivalents used to translate them: the most frequent equivalent in the training corpus is often the one that translates most of the instances of w in the test corpus.

The baseline score corresponds to both recall and precision, as a prediction is made for all the new instances. This score is calculated, for each w, on the basis of the number of its instances for which the proposed sense is correct. This number coincides with the frequency of the most frequent equivalent of w in the test corpus. In order to facilitate the comparison between our results and the baseline, we use the *f-measure* (*f-score*) that combines precision and recall in a unique measure:

$$f - score = \frac{2 * (precision * recall)}{precision + recall}$$

We evaluate here the performance of our WSD method on the 150 ambiguous nouns of our sample. We observe that the *f-score* of our method easily overcomes the results of the baseline.

baseline	51.42%
enriched *f*-score	76.99%

The difference between these scores indicates the positive impact of the clustering information on the WSD results. As the senses are situated at a higher level of abstraction, the correspondences with the reference are established at a more abstract level than that of exact unigram correspondences.

Our results can be compared to those obtained in the multilingual lexical sample tasks of Senseval and SemEval. This comparison seems interesting although these tasks concern words of different parts of speech (nouns, verbs and adjectives). The systems participating at the multilingual English–Hindi lexical sample task of Senseval-3 are all supervised and they all perform better than the baseline (Chklovski *et al.*, 2004). This is interpreted by the authors as an indication of the clarity of the sense distinctions performed using translations, which provide sufficient information for the training of supervised classifiers. The systems performed better on the sense-tagged part of the data, showing that sense information may be helpful for the task of targeted word translation. In the English–Chinese lexical sample task of SemEval the unsupervised systems perform

worse than the baseline, contrary to the supervised ones (Jin *et al*., 2007).

5 Capturing semantic similarity during translation

5.1 Lexical selection based on WSD

In the experiments reported here, *lexical selection* refers to the translation of ambiguous SL nouns in context and not to that of whole sentences. Lexical selection is thus considered as a *blank-filling* task (Vickrey *et al*., 2005): the equivalents translating the SL nouns in the TL sentences of the test TUs are automatically replaced by a blank which has to be filled by the WSD or the lexical selection method. We give an example of a test TU containing the noun *implication*.

b) "*Any change to the current situation must be preceded by a rigorous study of its various* ***implications****, with the objective always being to guarantee a high-quality public service and to retain the current public operators and existing jobs.*" / "*Επίσης, σε οποιαδήποτε μεταβολή της σημερινής κατάστασης θα πρέπει πάντα να προηγείται μία εμπεριστατωμένη μελέτη των διαφορετικών* **[…]** *έχοντας διαρκώς κατά νου το στόχο της διασφάλισης μίας ποιοτικής δημόσιας υπηρεσίας, της διατήρησης των σημερινών δημοσίων φορέων παροχής υπηρεσιών και της κατοχύρωσης των σημερινών θέσεων εργασίας.*"

If a one-equivalent cluster is selected by the WSD method, this equivalent is retained as the translation of the SL word (cf. (*a*), section 3). On the contrary, when a bigger sense cluster is proposed, the most adequate equivalent for the TL context has to be selected. This is done by the lexical selection method, which filters the cluster and fills the blank in the TL sentence with the best translation according to the TL context.

The cluster retained during WSD as describing the sense of *implication* in (b) is {*συνέπεια, επίπτωση*}. Most often the clustered equivalents are near-synonyms translating the same source sense, but almost never absolute synonyms interchangeable in all TL contexts. Consequently, the cluster can be filtered by considering their differences.

In order to judge the equivalents' adequacy in the new TL context, the lexical selection method compares information coming from this context to information learned during training. Given that the training was performed on a lemmatized and POS-tagged corpus, the new TL context must be lemmatized and POS-tagged as well, in order to retain only the lemmas of the content words[10].

The information acquired during training and exploited here concerns the context features that differentiate the equivalents in the TL, as shown by the semantic similarity calculation in the TL side of the training corpus (cf. section 2). The differentiating contexts of the equivalents characterize the sense clusters as well, as was the case with their assimilative contexts.[11]

The equivalent retained by the lexical selection method for *implication* in the example (b) is *συνέπεια*. This differs from the reference translation (*επίπτωση*) but is closely related to it. Thus, it is a semantically plausible translation that can be used in this TL context.

In a real Statistical Machine Translation (SMT) system, the clusters could be filtered by the language model, on the basis of word sequence probabilities in translations. In this way, the most probable translation in the TL context, among the semantically pertinent alternatives included in the cluster suggested during WSD, would be selected.

5.2 Evaluation of the lexical selection

The lexical selection method has been applied to the WSD results on our lexical sample. The reference translations, found in the test corpus, serve for evaluation here as well. We calculate the results of this method first using the principle of *strict precision* (i.e. looking for exact correspondences with the reference) and then on the basis of *enriched precision* (i.e. exploiting the clustering information).

The sense clusters serve here to estimate the semantic proximity of the proposed translation to the reference, in cases of no exact correspondence. Thus, a translation which is semantically similar to the reference is considered to be correct if they are both found in the cluster proposed during WSD. This renders the evaluation more flexible and significantly increases the quantity of semantically pertinent translations compared to the baseline.

The strict and enriched *f*-scores are estimated by considering as correct (score = 1) every translation that is pertinent according to the corresponding evaluation principles. The

[10] Our test corpus has been tagged and lemmatized using the TreeTagger (Schmid, 1994).

[11] The SL contextual information exploited for WSD.

results indicate the increase in pertinent translations.

baseline	52.14%
strict - *f-score*	48.37%
enriched *f-score*	77.79%

We observe that the strict *f-score* is lower than the baseline. This happens because our method proposes equivalents semantically similar to the reference for some instances for which the baseline predictions are correct. However, these pertinent predictions are not taken into account by the principle of strict precision. This is the case in example (b): the baseline prediction (*επίπτωση*) for this instance of *implication* corresponds to the reference while the suggestion of our method (*συνέπεια*), even though semantically pertinent, is not considered as correct according to the principle of strict precision and is not rewarded.

Nevertheless, it would be preferable to weigh differently the predictions related to the reference, by taking into account the strength of their relation. These predictions could be considered as *almost correct* and they could be, at the same time, penalized less than translations having a different sense and less rewarded than exact correspondences to the reference.

For this to be done, a measure capable of capturing the semantic distance would be needed. Using a weighted coefficient is essential in tasks implicating semantics, not only in WSD (Resnik and Yarowsky, 2000) but also in tasks such as the estimation of inter-annotator agreement in semantic annotation (Artstein and Poesio, 2008). The common element between these tasks is that the distances between the categories (word senses) should be weighted, so that the WSD errors or the divergences between annotators be treated differently.

We envisaged the possibility of weighting differently the proposed translations on the basis of their relation to the reference, by using as distance measure their similarity score in the TL. A semantically pertinent translation different from the reference was assigned a score equal to the similarity score of the two equivalents in the TL. A problem that we encountered, and that made us fall back to the solution of a uniform weighting of semantically pertinent translations, is that the comparison of these results to the baseline was not representative of the effective improvement (the great increase in the number of pertinent translation predictions) brought about by exploiting the clustering information. This happens because all the correct suggestions of the baseline are weighted by a score equal to 1, while the score of translations semantically related to the reference is always lower than 1, given that absolute synonyms are very rare in natural language.

We envisage the elaboration of a more sophisticated coefficient for weighting semantically pertinent translations, that will permit a more conclusive comparison with the baseline. This coefficient could take into account not only the similarity score between a proposed translation and the reference but also the number of the SL word's candidate translations, the number of its senses and their distinctiveness, as well as the number of the equivalents similar to the reference and their scores.

Before concluding, we would like to take a look at the way the concern for lexical semantics is manifested and taken into account in existing MT evaluation metrics.

5.3 Semantic similarity in existing MT evaluation metrics

Lexical semantic relations are supposed to be captured in BLEU by the use of multiple reference translations (Papineni *et al.,* 2002). Finding many references for evaluation is, however, rather problematic (Callison-Burch, 2006).

In METEOR (Banerjee and Lavie, 2005), such relations are detected by exploiting WordNet (Miller *et al.*, 1990). More precisely, the number of pertinent translations is increased using synset information: a translation is correct not only if it corresponds to the reference, but also if it is semantically similar to it, i.e. found in the same synset.

One of the limitations of this metric is that the words being tested for synonymy are not disambiguated; that is what Banerjee and Lavie call "a poor-man's synonymy detection algorithm". Consequently, the WN-Synonymy module used maps two unigrams together simply if at least one sense of each word belongs to the same WordNet synset.

Another problem is that the metric is strongly dependent on a predefined sense inventory. Given that such resources are publicly available for very few languages, the synonymy module often is not operational and is omitted.

Lavie and Agarwal (2007) envisage the possibility of developing new synonymy modules for languages other than English, which would be based on alternative methods and could replace WordNet.

In the previous sections, we showed how the information acquired by an unsupervised sense induction method can help to account for the words' semantic similarity. The created sense clusters, grouping semantically similar equivalents, can be compared to WordNet synsets. This kind of semantic information, extracted directly from text data, can constitute an alternative to the use of predefined sense inventories. A clear advantage of a metric based on the results of unsupervised semantic analysis, in comparison to one dependent on a predefined resource, is that it is language-independent and may be used for evaluation in languages where semantic resources are not available.

6 Conclusion and perspectives

In this paper, we have presented the advantages and weaknesses of cross-lingual sense determination, often used in multilingual WSD and MT. We have put forward some arguments towards a more thorough semantic analysis of the translation equivalents of ambiguous words that serve as sense indicators, and we have shown how it could be of use in multilingual WSD and MT.

The data-driven sense induction method used identifies the senses of ambiguous English nouns by clustering their translation equivalents according to their semantic similarity. Exploiting the sense inventory built in this way proves of benefit in multilingual WSD and lexical selection in MT. Their evaluation becomes more flexible as well, as it becomes possible to capture the semantic relations between the translations of ambiguous words.

The problem of strictness of the MT evaluation metrics can thus be overcome without the need for a predefined inventory. This would allow for a more conclusive estimation of the effect of WSD in SMT. The integration of the cluster-based WSD method into a real SMT system and the evaluation of its impact on translation quality constitute the main perspectives of the work presented in this article and the object of future work.

Acknowledgments

I would like to thank Philippe Langlais for the word alignment and Andy Way for useful comments. This research is funded by SFI grant 05/IN/1732.

References

Marianna Apidianaki. 2008. *Translation-oriented Sense Induction Based on Parallel Corpora*, In Proceedings of the 6th Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco.

Ron Artstein and Massimo Poesio. 2008. *Inter-coder Agreement for Computational Linguistics*, Computational Linguistics 34(4): 555-596.

Satanjeev Banerjee and Alon Lavie. 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*. In Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 43rd Annual Meeting of the Association for Computational Linguistics (ACL), Ann Arbor, Michigan, 65-72.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer. 1991. *Word-sense disambiguation using statistical methods*. In 29th Annual Meeting of the Association for Computational Linguistics (ACL), Berkeley, California, 264-270.

Clara Cabezas and Philip Resnik. 2005. *Using WSD Techniques for Lexical Selection in Statistical Machine Translation*. Technical Report CS-TR-4736/LAMP-TR-124/UMIACS-TR-2005-42.

Chris Callison-Burch, Miles Osborne and Philipp Koehn. 2006. *Re-evaluating the Role of BLEU in Machine Translation Research*. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento, Italy, 249-256.

Marine Carpuat and Dekai Wu. 2005. *Word Sense Disambiguation vs. Statistical Machine Translation*. In 43rd Annual Meeting of the Association for Computational Linguistics (ACL). Ann Arbor, Michigan, 387-394.

Marine Carpuat and Dekai Wu. 2007. *Improving Statistical Machine Translation using Word Sense Disambiguation*. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 61-72.

Yee Seng Chan, Hwee Tou Ng and David Chiang. 2007. *Word Sense Disambiguation Improves Statistical Machine Translation*. In 45th Annual

Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 33-40.

Timothy Chklovski, Rada Mihalcea, Ted Pedersen and Amruta Purandare. 2004. *The senseval-3 multilingual English-Hindi lexical sample task*. Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems, Barcelona, Spain, 5-8.

Philip Edmonds and Adam Kilgarriff. 2002. *Introduction to the special issue on evaluating word sense disambiguation systems*. Natural Language Engineering 8(4): 279-291.

Catherine, Fuchs. 1994. Paraphrase et énonciation. Editions Ophrys, Paris.

Maria Gavrilidou, Peny Labropoulou, Elina Desipri, Voula Giouli, Vasilis Antonopoulos and Stelios Piperidis. 2004. *Building parallel corpora for eContent professionals*. In Proceedings of the Workshop on Multilingual Linguistic Resources, 20th International Conference on Computational Linguistics (COLING), Geneva, Switzerland, 90-93.

Gregory Grefenstette. 1994. Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, Boston/Dordrecht/London.

Zellig Harris. 1954. Distributional Structure. *Word*, 10: 146-162.

Peng Jin, Yunfang Wu and Shiwen Yu. 2007. *SemEval-2007 Task 5: Multilingual Chinese-English Lexical Sample*, In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic, 19-23.

Philipp Koehn. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation*. In Proceedings of MT Summit X, Phuket, Thailand, 79-86.

Alon Lavie and Abhaya Agarwal. 2007. *METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments*. In Proceedings of the 2nd Workshop on Statistical Machine Translation, 45th Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic, 228-231.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Groos and Katherine Miller. 1990. *Introduction to WordNet: An On-line Lexical Database*. International Journal of Lexicography 3(4): 235-312.

George A. Miller and Walter G. Charles. 1991. *Contextual correlates of semantic similarity.* Language and Cognitive Processes 6(1): 1-28.

Karolina Owczarzak, Josef van Genabith and Andy Way. 2007. *Labelled Dependencies in Machine Translation Evaluation*. In Proceedings of the 2nd Workshop on Statistical Machine Translation, 45th Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic, 104-111.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA., 311-318.

Philip Resnik. 2004. *Exploiting Hidden Meanings: Using Bilingual Text for Monolingual Annotation*. In Gelbukh, A. (ed.), Lecture Notes in Computer Science 2945: Computational Linguistics and Intelligent Text Processing: Proceedings of the 5th International Conference CICLing, Seoul, Korea, 283-299.

Philip Resnik and David Yarowsk. 2000. *Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation*, Natural Language Engineering 5(3): 113-133.

Helmut Schmid. 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, 44-49.

David Vickrey, Luke Biewald, Marc Teyssier and Daphne Koller. 2005. *Word-Sense Disambiguation for Machine Translation*. In Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT-EMNLP), Vancouver, Canada, 771-778.