# Capturing Lexical Variation in MT Evaluation using Automatically Built Sense-cluster Inventories

Marianna Apidianaki, Yifan He, and Andy Way

NCLT/CNGL
School of Computing, Dublin City University
Dublin 9, Ireland
{mapidianaki, yhe, away}@computing.dcu.ie

**Abstract.** The strict character of most of the existing Machine Translation (MT) evaluation metrics does not permit them to capture lexical variation in translation. However, a central issue in MT evaluation is the high correlation that the metrics should have with human judgments of translation quality. In order to achieve a higher correlation, the identification of sense correspondences between the compared translations becomes really important. Given that most metrics are looking for exact correspondences, the evaluation results are often misleading concerning translation quality. Apart from that, existing metrics do not permit one to make a conclusive estimation of the impact of Word Sense Disambiguation techniques into MT systems.

In this paper, we show how information acquired by an unsupervised semantic analysis method can be used to render MT evaluation more sensitive to lexical semantics. The sense inventories built by this data-driven method are incorporated into METEOR: they replace WordNet for evaluation in English and render METEOR's synonymy module operable in French. The evaluation results demonstrate that the use of these inventories gives rise to an increase in the number of matches and the correlation with human judgments of translation quality, compared to precision-based metrics.

**Keywords:** Machine Translation, evaluation, synonymy, sense clustering

## 1 Introduction

The majority of the existing Machine Translation (MT) evaluation metrics look for exact surface correspondences between the compared translations. They are mostly based on the strict *precision* criterion, which does not account for the semantic similarity of words found in the hypothesis and the reference. Given that evaluation scores often do not reflect the quality of translation, there is a growing tendency towards increasing the correlation of the metrics with human judgments of translation quality. An important factor determining this correlation is the identification of sense correspondences between the hypothesis and the reference, which may exist even if the words used in the translations differ. Capturing this type of correspondence would also allow a more conclusive estimation of the impact of WSD techniques on MT systems than is possible with the current evaluation metrics (Callison-Burch *et al.*, 2006; Carpuat and Wu, 2005; Chan *et al.*, 2007).

In this paper, we show how variation at the unigram level can be captured during evaluation using information induced from parallel corpora by an unsupervised sense induction method. This method generates bilingual semantic inventories, where the senses of the words of one language are described by clusters of their semantically similar translation equivalents (TEs) in another language. These *sense-clusters*, which are similar to WordNet[1] synsets, can serve to capture correspondences between synonymous words found in the compared translations.

[1] http://wordnet.princeton.edu/

The structure of the paper is as follows: in the next section, we present how lexical variation is dealt with in existing MT evaluation metrics. In section 3, we describe the sense induction method and the training data used. In section 4, we explain how the automatically built sense inventory is integrated into METEOR. In section 5, we present the experiments carried out in English and in French and we analyze the obtained results. Then we show the advantages of integrating automatically acquired semantic information into MT evaluation. Finally we conclude, together with avenues for further work.

## 2 Lexical variation in existing evaluation metrics

BLEU (Papineni *et al.*, 2002) captures lexical variation by the use of multiple reference translations. However, this has been shown to be a rather problematic solution: even if numerous human translations of the same original text are available, which is rarely the case, their use poses additional problems during evaluation.[2]

METEOR (Banerjee and Lavie, 2005; Lavie and Agarwal, 2007) matches unigrams between the hypothesis and the reference in a flexible way, by using a *stemming* and a *synonymy* module. While the first matches different word forms, the second increases the number of pertinent translations by exploiting WordNet information: a translation is considered to be correct not only if it exactly corresponds to the reference, but also if it is semantically similar to it, i.e. found in the same WordNet synset. Nevertheless, predefined semantic resources like WordNet present some limitations. They cannot be easily updated and adapted to the domains of the processed texts and, most importantly, they are not publicly available for languages other than English. This is an important issue concerning METEOR as, when it is used for evaluation in languages other than English, only the *exact* and *stemming* matching modules are used, while the *synonymy* module is not operational and is omitted. This explains why Lavie and Agarwal (2007) propose to develop new synonymy modules for languages other than English, that would be based on alternative methods and could be used in the place of WordNet.

Some other metrics (Owczarzak *et al.*, 2007; He and Way, 2009) go beyond pure string matching. They look for “deeper” correspondences and thus correlate better with human judgments of translation quality. The above-mentioned metrics both use syntactic structure and dependency information in order to capture variations between sentences. In Owczarzak *et al.* (2007), lexical variation is also accommodated by adding WordNet synonyms into the matching process. In previous work by Owczarzak *et al.* (2006), lexical and syntactic paraphrases were extracted from the bitext used for evaluation using word and phrase alignment. In their work, the target language (TL) words/phrases aligned with each source language (SL) word/phrase constitute an *equivalence set*. The equivalents included in each set are considered to be synonyms or near-synonyms, in the case of words, and paraphrases, in the case of phrases.

An advantage of this procedure is that it allows the generation of paraphrases relevant to the domain of the text, which makes them more appropriate to the task at hand than synonyms extracted from an external resource like WordNet. However, an important issue concerning the equivalence sets is not addressed: the TL equivalents of a SL word/phrase are not always semantically related. SL words may be ambiguous, in which case their equivalents translate their different senses.

There are metrics that use paraphrases or textual entailment features to facilitate automatic evaluation, which are related to our proposed metric. Paraphrase-based metrics, such as ParaEval (Zhou *et al.*, 2006) and TERp (Snover *et al.*, 2009), use paraphrases mined from the corpus as “synonyms”; while in the Textual Entailment-based approach (Padó *et al.*, 2009), the metric tries to

[2] BLEU puts very few constraints on how *n*-gram matches can be drawn from the multiple reference translations and so it allows a too high amount of translation variation. Apart from that, the notion of *recall* - an important parameter in the evaluation of translation quality - is difficult to formulate over multiple reference translations and is not thus taken into account by BLEU (Callison-Burch *et al.*, 2006).

determine whether entailment can be inferred in both directions between hypothesis and reference. Our method differs from these approaches in the respect that it uses a different method to induce translation equivalents. Our method can also exploit the word sense information to create more fine-grained features for MT evaluation, which belongs to the avenues for future work.

In this paper, we show how the results of a data-driven semantic analysis method can be used in MT evaluation. This method induces the senses of SL ambiguous words from the data, by revealing the semantic relations of their TEs and distinguishing between semantically distant ones. We explain how the semantic information acquired by this method can be exploited by METEOR for evaluation. Exploiting this kind of information permits the capturing of domain-relevant synonymy relations, overrides the need for predefined resources and permits the use of METEOR's synonymy module for evaluation in languages other than English. The only requirement is that a parallel corpus, needed for training the sense induction method, be available in the concerned languages.

## 3 Data-driven sense induction

The semantic analysis method used here is the one proposed in Apidianaki (2008). Her method reveals the senses of ambiguous words of one language by clustering their TEs in another language. The created sense-clusters group semantically similar equivalents, whose relations are discovered from a parallel aligned training corpus. The method is based on the distributional hypotheses of meaning (Harris, 1954) and of semantic similarity (Miller and Charles, 1991), and on the assumption of sense correspondence between words in translation relation in real texts. The analysis is thus performed by combining distributional and translation information from a parallel corpus. Being totally data-driven this sense induction method is language-independent and permits the creation of sense inventories for different language pairs.

### 3.1 The training data

The training corpus used here is the sentence-aligned English(EN) – French(FR) part of Europarl (Koehn, 2005), which has been lemmatized and tagged by part-of-speech (POS) (Schmid, 1994). As the semantic analysis method is rather sensible to spurious alignments, a number of filters have been applied prior to word alignment in order to ensure the results with the least possible noise. First, function words were deleted in order to keep only the lemmas of content words. Then sentences containing more than five content words (and their translations) were deleted, as well as the sentence pairs presenting a great difference in length (cases where one sentence was three times longer than the other). After these filtering steps, word alignment was performed at the level of word types using Giza++ (Och and Ney, 2003).[3]

Two bilingual lexicons, one for each translation direction (EN–FR/FR–EN), were built from the alignment of word types. In these lexicons, each SL word (*w*) is associated with the set of TEs to which it was aligned.[4] Given the sensibility of the sense induction method to noise, some filtering steps have been applied at the level of the lexicons as well: first, the TEs of the ambiguous words were filtered on the basis of their score;[5] then, an intersection filter was applied, which discards any translation correspondences not found in both lexicons. While eliminating many false TEs, this process eliminated some good ones as well. The reason why we opted for this filtering is that the negative effect of the elimination of good TEs on the semantic analysis is less important than the noise present in the lexicons.[6]

[3] Aligning word types rather than tokens decreases data sparseness effects (Nießen and Ney, 2004).

[4] We aligned the corpus using two Giza++ configurations, with and without the *mkcls* component. As the lexicons generated from the two alignments contained some different entries (SL words), we kept their union in order to increase the coverage.

[5] The adopted threshold (0.03) was defined empirically.

[6] This could be described as an increase in *precision* – which is more important in lexicography applications (Och and Ney, 2003) – and a decrease in *recall*.

**Table 1:** Entries from the EN–FR and the FR–EN sense inventories.

| Language | POS | Source word | Sense-clusters |
|---|---|---|---|
| **EN–FR** | Nouns | disadvantage | {handicap},{désavantage, inconvénient} |
| | | shortcoming | {manquement}, {carence, insuffisance} |
| | Verbs | promote | {promouvoir},{favoriser, encourager} |
| | | reiterate | {réaffirmer},{répéter, réitérer} |
| | Adjectives | intolerable | {insupportable, intolérable},{intenable} |
| | | workable | {réalisable, viable}, {fonctionnel} |
| **FR–EN** | Nouns | répercussion | {repercussion, impact},{implication} |
| | | enseignement | {education}, {lesson, teaching} |
| | Verbs | insister | {highlight},{stress, insist, emphasise} |
| | | diffuser | {circulate, publicise}, {spread, broadcast, disseminate} |
| | Adjectives | rigoureux | {tough}, {strict, stringent, rigorous} |
| | | définitif | {final, definitive}, {permanent} |

A POS filter was then applied to the lexicons, in order to keep for each SL word *w* its TEs which pertain to the same category.[7] Given that the TEs of each *w* help to analyse its semantics, this filter serves an additional goal apart from eliminating spurious alignments: it facilitates the semantic analysis of the SL *w*s by resolving their eventual categorial ambiguity. Finally, all SL words having multiple (more than two) TEs were kept.

## 3.2 Unsupervised sense induction

The core component of the sense induction method used is a semantic similarity calculation which reveals the relations between the TEs of each SL word *w*. We call a *translation unit* (TU) a pair of aligned sentences. Each TU may contain up to two sentences of each language, which are in translation relation in the parallel corpus. A sub-corpus is created from the training corpus for each *w*, grouping the TUs where *w* appears in the SL sentence(s). These TUs are then grouped by reference to *w*'s TEs. So, if *w* has three TEs (*a*, *b* and *c*), three sets of TUs are formed (one where *w* is translated by *a* ('*w*-*a*' TUs), one where it is translated by *b* ('*w*-*b*' TUs), etc.).

We consider that each TE is characterized by a set of SL contextual features: the set of lemmatized content words surrounding *w* when it is translated by that TE.[8] These features are extracted and treated as a *bag of words*. This distributional information serves to estimate the TEs' similarity using a variation of the Weighted Jaccard coefficient (Grefenstette, 1994) described in Apidianaki (2008).

The TEs are compared in a pairwise manner and a similarity score is assigned to each pair. Two TEs are considered to be 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 their similarity score to a threshold, equal to the mean of the scores assigned to all the pairs of TEs of *w*. The similarity calculation results are exploited by a clustering algorithm which groups the TEs into clusters [9]. The generated clusters illustrate the senses of *w*: clustered TEs are semantically related and considered as translating the same SL sense, while isolated ones translate distinct senses.

# 4 Integrating automatically acquired semantic information into METEOR

The sense induction method described above permits the automatic creation of a *sense-cluster* inventory for each of the languages represented in the parallel corpus. Two such inventories have

[7] The noun equivalents of nouns, the verb equivalents of verbs, etc. This filtering is performed by using a POS lexicon built from the tagged training corpus, where every word of each language is assigned its possible tags.

[8] For instance, *a* is characterized by the words surrounding *w* in the SL sentences of the *w*-*a* TUs set.

[9] The properties of the clustering algorithm are described in Apidianaki (2008).

been created from our training corpus: a EN–FR one (where the senses of EN ambiguous words are described by clusters of their FR TEs) and a FR–EN one (where the senses of FR words are described by clusters of their EN TEs). These inventories contain information on Nouns, Verbs and Adjectives. In Table 1, we give some entries from the two inventories for words of different POS categories. The clusters, found in the fourth column, describe the senses of the words found in the third column. These sense-clusters group semantically similar words and are thus similar to WordNet synsets.

We propose to replace and compare with the WN module in METEOR for two reasons. Firstly, as the WN module of METEOR is shown to improve the correlation with human judgment, it would be interesting to see if our method has the same effect. Secondly, METEOR is a well established and stable metric and improvement on such a metric would be of more practical use.

A clear advantage of this automatically created inventory in comparison to WordNet is that the information is acquired directly from corpora. It is thus relative to the domains of the processed texts and may concern languages for which WordNet-type resources are not available. Additionally, Apidianaki (2009) showed that this data-driven sense induction method provides, as a by-product, information that can be exploited by an unsupervised Word Sense Disambiguation classifier. It consists of the SL distributional information that reveals the similarity of the TEs and can serve to disambiguate new instances of the ambiguous words. This is another advantage of this unsupervised method: it makes it possible to carry out a disambiguation step during evaluation, which could replace METEOR's "poor-man's synonymy detection algorithm" (Banerjee and Lavie, 2005).

Nevertheless, this inventory is also characterised by the weaknesses of automatically built resources. On the one hand, the filterings applied to the training corpus (cf. section 3.1) do not manage to eliminate all the noise found in the word-alignment results, partly due to POS-tagging and lemmatization errors. On the other hand, these filterings eliminate pertinent translation information that is thus missing from the final resource.

**Table 2:** French translations with sense correspondances.

| **Reference** | **Hypothesis** |
|---|---|
| Le Parlement européen n'a qu'une façon de **prouver** qu'il se respecte ... | Le Parlement européen vient d'une façon de **montrer** qu'elle a le respect ... |
| ... il ne serait, je pense, pas dans leur intérêt, ou dans l'intérêt des États membres, d'**atteindre** ce stade sans ... | ... il ne serait pas, je crois, être dans leur meilleur intérêt, voire les intérêts des États membres, afin de **parvenir** à cette échéance sans ... |
| Pour cela, il est absolument **indispensable** d'élaborer des programmes de ... | Pour que cela soit le cas, il est absolument **essentiel** d'élaborer des plans pour ... |
| Je peux vous **assurer**, Mesdames et Messieurs les Députés, que c'est une question dont je m'occuperai ... | Je peux **garantir** aux honorables députés que c'est une question que j'aborderai ... |
| Quoi qu'il en soit, l'engrenage mis en place fait son œuvre et nous imposera sa logique jusqu'au **moment** où ... | Néanmoins, le système qui a été mis à la place fera son travail et nous imposera sa logique jusqu'à ce que le **temps** vienne quand ... |

However, we consider that this automatically elaborated resource can be exploited for identifying word matches in languages other than English. The sense-cluster information can be used to account for lexical variation in translation, by revealing the semantic relations that may exist between the words found in the hypothesis and the reference.

In this work, the inventory acquired for French is exploited for rendering METEOR's synonymy module operable for MT evaluation in French. Additionally, a comparison is made between the results obtained when the automatically built English inventory replaces WordNet into METEOR. In what follows, we describe the experiments carried out in both languages and show how exploiting this information renders the evaluation more sensitive to lexical semantics and closer to human judgments of translation quality.

## 5 Experiments

### 5.1 Evaluation in French

**MT evaluation** We evaluated the baseline system of the WMT08 English–French shared task (Moses, Koehn *et al.*, 2007) using METEOR, with and without exploiting the cluster inventory (i.e. performing only *exact* and *porter_stemmer* match). The results are given in Table 3.

**Table 3:** Baseline results on the WMT08 EN–FR data.

| | **METEOR** | **METEOR_syn** |
|---|---|---|
| **Matches** | 18225 | **18693** |
| **Chunks** | 8558 | **8949** |
| **Final Score** | 0.1829 | **0.1836** |

We observe that the numbers of matches (original score) and chunks (penalty) and the final score all increase when the cluster inventory is used. It is important to note that the chunk penalty in METEOR has a higher weight for French (1) than for English (0.28), so for English, the increase in final score should be bigger. The results obtained for English are presented in section 5.2.

In Table 2 we present some cases where METEOR manages to identify sense correspondences between the hypothesis and the reference using the cluster information, that would otherwise be missed.

**Correlation with Human Judgments** In order to calculate the correlation that METEOR has with human judgments during evaluation in French, we use the WMT08 evaluation shared task dataset.[10] All English–French human rankings (307 in total), distributed during this shared evaluation task for estimating the correlation of automatic metrics to human judgments of translation quality, were used for our experiments. The rankings provided here are at the level of the segment.

To measure the *correlation* of the automatic metrics with the human judgments of translation quality, we use Spearman's rank order correlation coefficient (Callison-Burch *et al.*, 2008). Spearman's correlation is defined as in (1), where $d$ is the difference between corresponding values in rankings and *n* is the length of the rankings.

$$\rho = 1 - (\frac{6 \sum d^2}{n(n^2 - 1)}) \tag{1}$$

An automatic evaluation metric with a higher correlation value is considered to make predictions that are more similar to the human judgments than a metric with a lower value.

For measuring the *consistency* of the automatic metrics with human judgments, we use the *pairwise consistent percentage* (Callison-Burch *et al.*, 2008). For every pairwise comparison of two systems on a single sentence by a person, the automatic metric is counted as being consistent if the relative scores are the same (i.e. the metric assigned a higher score to the higher ranked system). The pairwise consistent percentage is equal to the number of correct pairwise comparisons made by a metric divided by the total number of pairwise comparisons performed.

In Table 4, we present the (sentence-level) results of the correlation of METEOR with human judgments, with and without the synonymy module.

[10] http://www.statmt.org/wmt08/shared-evaluation-task.html

**Table 4:** Correlation results in French.

| | **BLEU** | **Meteor** | **Meteor_syn** |
|---|---|---|---|
| Spearman.Cor. | 0.2078 | 0.2657 | **0.2687** |
| Consistency | 0.4571 | **0.6279** | 0.6277 |

In the consistency test, the two metrics are both required to perform 967 pair-wise predictions. Of these predictions, METEOR without the synonymy module makes two more correct predictions while the two metrics make 13 different predictions. However, METEOR with synonymy has a better overall correlation with human judgments. Because of the small amount of human judgments data available for French, we expect correlation results in English to be more indicative of the benefit of integrating the sense-cluster inventory into METEOR.

## 5.2 Evaluation in English

**MT evaluation** For our English experiments, we first used the WMT08 French-English data set. The results obtained by METEOR on this data set are given in Table 5. We compare three test settings: METEOR without synonymy (*plain*), with WordNet synonymy (*wn*) and with sense-cluster synonymy (*wsd*).

**Table 5:** Results on the FR–EN data from WMT08.

| | **plain** | **wn** | **wsd** |
|---|---|---|---|
| **Matches** | 134620 | **144788** | 140356 |
| **Chunks** | 78892 | **84356** | 82703 |
| **Score** | 0.4623 | **0.4977** | 0.4815 |
| **Correl.** | 0.2949 | **0.3062** | 0.2997 |
| **Consist.** | 0.6173 | **0.6262** | 0.6199 |

According to these results, when using the sense-cluster inventory METEOR finds 4.09% more matches than *plain*, and 3.06% fewer matches than when WordNet is used. As we expected, the increase in the final score is bigger than in the case of French when the cluster inventory is used. Concerning the correlation of METEOR with human judgments, we observe an increase in both *correlation* and *consistency* when the cluster inventory is used, in comparison to *plain*.

We also conducted experiments on the MTC4 corpus from LDC, which consists of human-assigned *fluency* and *adequacy* scores to 11,028 sentences generated by 11 MT systems. The results obtained on the MTC4 data set are given in Table 6. We calculate Pearson's correlation with human judgments on both fluency and adequacy. Pearson's correlation is defined as in (2):

$$r = \frac{1}{n-1}\sum(\frac{x_i - \bar{X}}{s_X})(\frac{y_i - \bar{Y}}{s_Y}) \qquad (2)$$

where $x_i$ is the value of the $i^{th}$ score, $\bar{X}$ is the mean score and $s_X$ is the standard deviation. According to these results, when using the sense-cluster inventory, METEOR finds 4.25% more matches than *plain* and 4.4% fewer matches than WordNet. Regarding correlation, we observe a descrease in adequacy and an increase in fluency in comparison to *plain*.

**Comments on the results** Some interesting comments can be made on these results. First of all, in order to interpret them correctly, we have to take into account the coverage of the resources that are being used.

On one hand, our English sense inventory counts 2,078 sense-clusters, from which only 1,555 contain more than one word. The total number of elements in the clusters is 4,004, while the

**Table 6:** Results on the FR–EN data from MTC4.

| | **plain** | **wn** | **wsd** |
|---|---|---|---|
| **Matches** | 148633 | **162377** | 155231 |
| **Chunks** | 104600 | **113619** | 110074 |
| **Score** | 0.3835 | **0.4195** | 0.3998 |
| **Adequacy** | 0.3253 | **0.3373** | 0.3248 |
| **Fluency** | 0.1699 | **0.1718** | 0.1713 |

average cluster size is 1.92 elements. If we consider only the clusters with more than one element, the average cluster size is 2.23 elements. On the other hand, WordNet counts 664,679 synsets, from which 345,501 contain more than one word. The average synset size is 1.96 while, when considering only the synsets with more than one element, it is 2.84.

Looking at these statistics, it seems normal that WordNet's coverage is greater than that of our inventory. The number of the WordNet synsets containing more than one word (345,501) is much higher than the corresponding number of clusters (1,555). Apart from that, the average size of the WordNet synsets that contain more than one word (2.84) is bigger than that of the corresponding sense-clusters (2.23). These quantitative differences explain to a great extent why WordNet manages to find more matches.

At the same time, the fact that METEOR using the cluster inventory finds much more matches than the *plain* METEOR configuration and only 3.06% fewer than METEOR with WordNet on the first test set (WMT08), while it is situated half way between *plain* and METEOR with WordNet on the second test set (MTC4), is rather promising. This means that if the inventory contained more clusters and clusters with a larger number of elements, its coverage would get closer to that of WordNet. The small coverage of the resource used here is due to its fully automatic elaboration. As we have already noted, the number of TEs found in the automatically generated lexicons is rather small, due to the subsequent filtering of the word-alignment results.

However, what the evaluation results also show is that a small inventory, automatically created from a parallel corpus, can cover half of the cases taken into account by an expensive resource like WordNet during MT evaluation. Even if more work has to be done in order to increase its coverage, it still manages to capture a great deal of the matches between the hypotheses and the references.

## 6 Conclusion and perspectives

In this paper, we have shown how a semantic resource automatically generated from a parallel corpus can be exploited in MT evaluation for capturing lexical variation. This resource contains clusters of semantically related translation equivalents of SL words, which can be used to detect synonymy and capture lexical variation during evaluation.

The cluster inventory created for French has been integrated into an existing automatic MT evaluation metric, METEOR. The exploitation of this resource renders operable in French the *synonymy* module of METEOR, which would otherwise be omitted. The evaluation performed in French shows that more matches are found with this configuration of METEOR, than when only the *exact* and *stemming* modules are used. Given that the sense-induction method used is language-independent, the same could be done in other languages not disposing of WordNet-type resources for rendering the evaluation more sensitive to lexical semantics. The only requirement would be that a parallel corpus be available in the languages concerned, in order for the semantic analysis method to be able to generate the sense inventory needed.

The merit of this approach is shown during evaluation in English, as well. Even if fewer matches are found between the hypothesis and the reference when the cluster inventory is used compared to when WordNet is used, an increase in the number of matches is observed relatively to when only the exact and stemming modules are used. These results are encouraging and point to perspectives for increasing the resource's coverage. This could be done by using more training corpora but, most importantly, by improving the quantity of the information found in the lexicons generated from the word-alignment results. Alternative ways of eliminating the noise found in these results should be investigated in order to avoid the need for numerous filtering steps, which decrease the information found in the translation lexicons.

Another point that is worth noting here is that a resource much smaller than WordNet[11] manages to capture half of the matches found by WordNet. A conclusion that could be drawn from this is that WordNet effectively contains a great amount of information that is really irrelevant for MT evaluation in specific domains. This demonstrates once again the importance of using resources directly generated from data for evaluation. We consider that it would be interesting to proceed to a more thorough analysis of the matching results, in order to measure the complementarity of these resources and of the cases they cover. This could permit the drawing of more pertinent conclusions on this aspect and would also allow us to estimate the benefit of enriching hand-crafted resources with automatically acquired information. As part of future work, we intend to generate sense-cluster inventories from different data sets and for more language pairs, and use them in MT evaluation.

## References

Apidianaki, M. 2008. Translation-oriented Sense Induction Based on Parallel Corpora. *Proceedings of the 6th Conference on Language Resources and Evaluation (LREC)*, pp. 3269-3275. Marrakech, Morocco.

Apidianaki, M. 2009. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 77-85. Athens, Greece.

Banerjee, S. and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL-2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 65-72. Ann Arbor, Michigan.

Carpuat, M. and D. Wu. 2005. Word sense disambiguation vs. statistical machine translation. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 387-394. Ann Arbor, Michigan.

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

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz and J. Schroeder. 2008. Further Meta-Evaluation of Machine Translation. *Proceedings of the ACL-2008 Workshop on Statistical Machine Translation*, pp. 70-106. Columbus, OH.

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

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

[11] WordNet is 222 times bigger than the sense-cluster inventory, if we count only the clusters and the synsets with more than one words.

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

He, Y. and A. Way. 2009. Learning Labelled Dependencies in Machine Translation Evaluation. *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT)*, pp. 44-51. Barcelona, Spain.

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

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the ACL, Demo and Poster Sessions*, pp. 177-180. Prague, Czech Republic.

Lavie, A. and A. Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation*, pp. 228-231. Prague, Czech Republic.

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

Nießen, S. and H. Ney. 2004. Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics*, 30(2), 181-204.

Och, F.J. and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1), 19-51.

Owczarzak, K., D. Groves, J. van Genabith and A. Way. 2006. Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. *Proceedings of the HLT-NAACL 2006 Workshop on Statistical Machine Translation*, pp. 86-93. New York, USA.

Owczarzak, K., J. van Genabith and A. Way. 2007. Labelled Dependencies in Machine Translation Evaluation. *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation*, pp. 104-111. Prague, Czech Republic.

Padó, S., M. Galley, D. Jurafsky and C. Manning. 2009. Robust Machine Translation Evaluation with Entailment Features. *Proceedings of ACL-IJCNLP*, pp. 297-305. Suntec, Singapore.

Papineni, K., S. Roukos, T. Ward and W.-J. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311-318. Philadelphia, PA.

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

Snover, M., N. Madnani, B.J. Dorr and R. Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. *Proceedings of the EACL-2009 Workshop on Statistical Machine Translation (WMT09)*, pp. 259-268. Athens, Greece.

Zhou, L., C.-Y. Lin and E. Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 77-84. Sydney, Australia.