An algorithm for cross-lingual sense-clustering tested in a MT evaluation setting

Marianna Apidianaki

Alpage, INRIA & Université Paris 7
175 rue du Chevaleret
Paris 75013, France
marianna.apidianaki@inria.fr

Yifan He

CNGL, School of Computing
Dublin City University
Dublin 9, Ireland
yhe@computing.dcu.ie

Abstract

Unsupervised sense induction methods offer a solution to the problem of scarcity of semantic resources. These methods automatically extract semantic information from textual data and create resources adapted to specific applications and domains of interest. In this paper, we present a clustering algorithm for cross-lingual sense induction which generates bilingual semantic inventories from parallel corpora. We describe the clustering procedure and the obtained resources. We then proceed to a large-scale evaluation by integrating the resources into a Machine Translation (MT) metric (METEOR). We show that the use of the data-driven sense-cluster inventories leads to better correlation with human judgments of translation quality, compared to precision-based metrics, and to improvements similar to those obtained when a handcrafted semantic resource is used.

1. Introduction

Electronic semantic resources are available in a few languages and are often criticized as being inappropriate for automatic processing. This is mainly due to their limited coverage, as well as to the too fine granularity of the proposed semantic descriptions. Even though resources of this kind are required for Word Sense Disambiguation (WSD), it has been shown that fine distinctions are rarely needed and that the majority of the applications just need coarse-grained WSD [1, 2, 3].

However, semantic resources are also needed for evaluation purposes. Information on word relations and sense proximity is necessary for a differing penalization of WSD errors [4] and enhances the pertinence of Machine Translation (MT) evaluation metrics [5]. Nevertheless, the majority of the existing metrics are looking for exact surface correspondences while attempts to capture semantic relations are hampered by the scarcity of large-scale resources.[1] Consequently, there is a growing interest for developing synonymy modules based on alternative methods [7]. On the basis of these observations, [8] exploited small data-driven semantic resources in MT evaluation and reported slight improvements compared to precision-based metrics.

In this paper, we present a clustering algorithm that creates large-scale bilingual semantic resources from parallel corpora. In these resources, the senses of the words of one language are described by clusters of their semantically similar translations in another language. We show that the integration of these *sense-cluster* inventories into a MT evaluation metric (METEOR) leads to important improvements : a better correlation of the metric with human judgments of translation quality, compared to precision-based metrics, and a performance similar to the one observed when a large handcrafted resource (WordNet) is used.

The paper is structured as follows: in the next section, we present the main approaches to cross-lingual sense induction and the method used here. In section 3, we describe the training data and the generated bilingual resources. In sections 4 and 5, we present and analyze the results of a large-scale extrinsic evaluation performed by integrating the resources into a MT evaluation metric. Finally, we conclude, together with avenues for future work.

2. Unsupervised sense-induction

2.1. Translations as sense indicators

Unsupervised sense-induction methods identify the senses of polysemous words in texts, without need for classified training data. In a monolingual context, this is done by *clustering* the instances of polysemous words in text corpora according to their distributional similarity [9, 10, 11]. In a bi-lingual (and multi-lingual) context, the senses of polysemous words are usually identified by considering their translation equivalents (TEs) as sense indicators [12, 13, 14]. These sense-induction methods are based on the assumption that the TEs of the words, found in the word alignment results of parallel corpora, lexicalize different source language (SL) senses in the target language (TL).

This approach permits to adapt the sense descriptions to the needs of multilingual applications. Additionally, it over-

[1] For instance, the *synonymy* module of METEOR exploits WordNet [6], which restricts its application to one language.

rides the need for predefined sense inventories and puts the accent on semantic distinctions pertinent for the concerned language pairs [4]. However, the possible semantic relations between the TEs of a SL word are not taken into account. Moreover, the TEs may reflect semantic distinctions pertinent only in the TL or present the same ambiguity as the SL word. Not considering these phenomena results in the induction of uniform senses, which raises problems in multilingual WSD and its evaluation [15].

Nevertheless, clustering techniques can also be used for sense induction in a multi-lingual setting. One possibility is to build vectors for the instances of polysemous words in a multilingual parallel corpus by using as features the TEs to which they are aligned [16]. By clustering the vectors according to their similarity the corresponding instances are grouped as well, and the obtained clusters describe the senses of the words. Another possibility, when working with parallel corpora, is to perform a clustering in the TL by using SL feature vectors. In the method proposed by [17], each TE of a SL word *w* is characterized by a vector built from the SL cooccurrences of *w*, whenever it is translated by this TE in the aligned sentences of the parallel training corpus. The TEs are then grouped into clusters on the basis of the similarity of their feature vectors. The obtained *sense-clusters* describe the senses of the SL word. An important merit of this method is that the semantically similar TEs are grouped and are not considered as indicators of distinct senses. Furthermore, word-sense relations are identified and described by the overlaps between the resulting clusters.

In this work, we build bilingual semantic inventories by using a modified version of this sense-induction method. The clustering algorithm has been optimized and the procedure that defines the threshold used for clustering is different than the one used in the original method, and better adapted to the data. The process employed for building the bilingual lexicons needed for clustering has been modified as well, as we will show in section 3, in order to increase the coverage of the obtained resources.

2.2. Clustering for cross-lingual sense induction

2.2.1. Semantic similarity

The parallel corpus used for training our cross-lingual sense induction method has to be lemmatized, tagged by part-of-speech (POS), and aligned at the sentence and word levels. In order to build the feature vectors of the TEs of a polysemous SL word *w*, we retain the lemmas of the content words (Nouns, Verbs and Adjectives) that cooccur with *w* in the aligned sentences where it is translated by each of the TEs. The similarity of the feature vectors is calculated by using a variation of the Weighted Jaccard measure [17, 18]. This metric weighs each SL feature according to its relevance for the estimation of the similarity of the TEs. The pairwise similarity of the TEs of a *w* is then estimated by comparing the corresponding weighted SL feature vectors. A similarity score is assigned to each pair of TEs (TE_pair) and stored in a table that is being looked up by the clustering algorithm. So, the similarity calculation does not have to be repeated during clustering.

2.2.2. Dynamic thresholding

The pertinence of the relation of each TE_pair is estimated by comparing its similarity score to a threshold. The threshold is defined locally for each *w* by the following iterative method.

1. The initial threshold (T) corresponds to the mean of the scores (above 0) of the TE_pairs of *w*.
2. The set of TE_pairs is segmented into pairs whose score exceeds the threshold and pairs whose score is inferior to the threshold, creating two sets (G1, G2).
3. The average of each set is computed (*m1* = average value of G1, *m2* = average value of G2).
4. A new threshold is created that is the average of *m1* and *m2* (T = (m1 + m2)/2).
5. Go back to step 2, now using the new threshold computed in step 4, keep repeating until convergence has been reached.

Experimentally this threshold adapts better to the data and gives better results than the mean of the similarity scores, used in [17].

2.2.3. The SEMCLU algorithm

The SEMCLU (SEMantic CLUstering) algorithm builds the clusters of TEs describing the senses of a *w* by exploiting the similarity calculation results. The input of SEMCLU for each *w* consists in : (a) the list of its TEs (TE_list); (b) their similarity table; (c) the similarity threshold.

The clustering is performed in two steps. First, each TE_pair, having a similarity score above the threshold, is considered to have a pertinent relation and forms an 'initial cluster' (C). These two-element clusters are derived directly from the similarity table. During the second step, the two-element clusters may be enriched by additional TEs, by the recursive function 'enrich_cluster' shown in Algorithm 1. The function takes as input the cluster C and the list of TEs of *w* (TE_list), and outputs C eventually enriched by other TEs.

The condition for a new TE to be included in C is to have a pertinent relation with all the elements already in C. The clustering stops when all the TEs of *w* are included in some cluster and all their relations have been checked. The final clusters are characterized by *global connectivity*, i.e. all their elements are linked by pertinent relations. The TEs having no pertinent relation to any other TE of *w* are included in separate one-element clusters. In the next section, we present the semantic resources created by this clustering algorithm.

Language	POS	Source word	Sense-clusters
EN–FR	Nouns	restriction	{interdiction}{restriction,limitation}{contrainte,limitation}{réduction}{entrave}
		maintenance	{entretien,maintenance} {entretien,maintien} {préservation}
	Verbs	accommodate	{adapter,répondre} {accueillir} {satisfaire,répondre}
		combine	{conjuguer,combiner,associer} {réunir,unir,conjuguer,concilier} {conjuguer, concilier,réunir,associer} {regrouper,rassembler,réunir} {ajouter} {fusionner}
	Adjectives	thorough	{profond,complet}{exhaustif,sérieux,minutieux}{profond,sérieux}{rigoureux} {complet,minutieux}
		tough	{dur,âpre} {dur,sévère,ferme,strict} {rude} {fort}
FR–EN	Nouns	limitation	{restriction,limitation,limit} {reduction}
		force	{strength,force} {violence} {strength,power}
	Verbs	illustrer	{illustrate} {demonstrate,show,exemplify} {reflect,exemplify,show}
		aider	{assist,aid,support} {enable} {aid,enable,assist} {contribute}
	Adjectives	contradictoire	{contradictory, conflicting} {inconsistent}
		épineux	{tricky, difficult} {tricky, sensitive} {thorny, difficult}

Table 1: *Entries from the sense-cluster inventories*.

Algorithm 1 *The 'enrich_cluster' function.*

```
enrich_cluster(TE_list, C):
if empty TE_list then
  return C
else
  if first TE in TE_list linked to all TEs in C then
    enrich_cluster(rest TE_list, C union TE)
  else
    enrich_cluster(rest TE_list, C)
  end if
end if
```

3. Bilingual sense-cluster inventories

3.1. The training corpus

The parallel corpus used here for training the sense induction method is the sentence aligned French (FR) – English (EN) part of Europarl (release v5, containing 1,723,705 sentence pairs) [19]. The corpus is lemmatized, POS tagged [20] and aligned at the level of word types using Giza++ [21]. The sentence pairs presenting a great difference in length (cases where one sentence is three times longer than the other) are eliminated.[2]

Two bilingual lexicons (one for each translation direction : EN–FR / FR–EN) are built from the alignment of word types, where each *w* is associated with the TEs to which it is aligned. The lexicons contain only content words (Nouns, Verbs and Adjectives). Some of the proposed TEs are filtered on the basis of their score.[3] Then an intersection filter is applied, which discards any translation correspondences not found in both lexicons. Finally, the two lexicons are filtered by POS, keeping for each *w* only its TEs pertaining to the same category.

The TEs used for clustering are the ones that translate *w* more than 10 times in the training corpus. We noticed that when a high number of equivalents (over 20) is retained for a *w*, this is a very frequent word which is aligned to target language words that are not good equivalents. In these cases, we increase the frequency threshold and we only retain the equivalents that translate *w* over 30 times in the corpus. Even if these thresholds leave out some TEs of the source words, they have a double merit : they reduce data sparseness issues and eliminate erroneous TEs which may be found in the lexicons because of spurious alignments.

3.2. The bilingual resources

Two resources have been generated by the sense induction method from our training data : a EN–FR inventory, where the senses of EN words are described by clusters of their TEs in French; a FR–EN inventory, where the senses of FR words are described by clusters of their semantically similar TEs in English. The EN–FR inventory contains 3,737 English words (Nouns : 2,166, Adjectives : 581, Verbs : 990) and 13,388 clusters.[4] The clusters contain in average 2.66 elements. The FR–EN inventory contains 3,734 French words (Nouns : 2,190, Adjectives : 572, Verbs : 972) and 11,775 clusters with 4.07 elements in average.[5]

The sense-clusters group semantically similar translations of the source words and could be compared to WordNet synsets. Each sense-cluster describes a sense of the source word. In Table 1, we present some examples of English and French words of different parts-of-speech whose senses are described by clusters of their TEs. The English word *maintenance*, for instance, has four TEs (*entretien, maintenance,*

[2] Aiming to obtain large coverage sense-cluster inventories, we did not apply the other filters described in [8] prior to word alignment.

[3] The low threshold (0.001) ensures that many alignment correspondences are kept.

[4] 4,699 clusters with 1 TE, 4,319 with 2, 1,572 with 3 and 2,798 clusters with >3 French TEs.

[5] 1,605 clusters with 1 TE, 3,734 with 2, 1,524 with 3 and 4,912 with >3 English TEs.

préservation, maintien) which are grouped in three clusters describing its senses: {*entretien, maintenance*}, {*entretien, maintien*} and {*préservation*}. We observe that the first two clusters overlap (they both contain the word *entretien*), which means that the two senses of *maintenance* are related. In the case of the French word *limitation*, its TEs are grouped into two disjoint sense-clusters : {*restriction, limitation, limit*} and {*reduction*}, describing two distinct senses of the word.

4. Evaluation

4.1. Intrinsic vs extrinsic evaluation

Automatically generated resources can be evaluated *intrinsically*, against a gold standard, or *extrinsically*, by measuring their contribution to the performance of some task. An intrinsic evaluation in lexical semantics is difficult because of the lack of a gold standard and, also, because the contents of existing semantic resources have been shown to be inappropriate for WSD in Natural Language Processing applications [2, 3]. Additionally, the WSD needs of the applications vary and this has an impact on the semantic resources that should be used. So, the results of an intrinsic evaluation of a resource would not be meaningful for its usefulness in different settings.

Another important factor in favor of an extrinsic evaluation of data-driven semantic resources is that unsupervised semantic analysis methods are often application-oriented. The sense-cluster inventories generated by the cross-lingual sense induction method used here are adequate for multilingual processing. Consequently, we choose to perform a large-scale extrinsic evaluation of the generated resources by integrating them in a MT evaluation metric and estimating the impact of their use during MT evaluation.

4.2. Exploiting sense-clusters in MT evaluation

4.2.1. Lexical variation in MT

MT evaluation results often do not reflect the quality of the translations provided by the MT systems. This is partly due to the fact that the majority of the existing MT evaluation metrics are based on the strict *precision* criterion, according to which exact surface correspondences between the compared translations have to be found. However, 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 *lexical variation* between the hypothesis and the reference, i.e. of sense correspondences which may exist even if the words in the translation differ.

Attempts to capture lexical variation concern : the use of multiple reference translations [22]; the identification of "deeper" correspondences between sentences by using syntactic structure and dependency information [23]; the detection of paraphrases [24, 25] or of textual entailment [26] between hypothesis and reference. METEOR [5, 7] matches unigrams between the compared translations 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.

The synonymy module of METEOR is shown to improve the correlation of the metric with human judgments of translation quality. Nevertheless, given that large-scale resources, like WordNet, are not publicly available for languages other than English, when METEOR is used for evaluation in other languages, only the *exact* and *stemming* matching modules are used, while the *synonymy* module is omitted.

4.2.2. Integrating sense-clusters into METEOR

In this work, we integrate our data-driven semantic inventories into METEOR. Given that the sense-clusters generated by SEMCLU are similar to WordNet synsets (i.e. they group semantically related words), it would be interesting to see if exploiting the inventories generated by our method has the same positive effect during evaluation. Like WordNet, the sense-cluster inventories 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. It is important to note that, unlike previous versions, METEOR version 1.0, used here, uses a built-in lemmatizer in order to enhance the identification of synonym matches between the hypothesis and the reference. However, the synonymy module is still available only for evaluation in English.

In the evaluation carried out for English, a comparison is made between the results obtained when the automatically built English inventory replaces WordNet into METEOR. The sense-cluster inventory acquired for French is exploited for rendering METEOR's synonymy module operable for MT evaluation in French. 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.

4.3. Evaluation in French

4.3.1. MT evaluation

We evaluated submissions to the WMT09 English–French shared task[6] using METEOR, with and without exploiting the sense-cluster inventory (i.e. performing only *exact* and *porter_stemmer* match). The results are given in Table 2.

We observe that the number of matches (original score) and chunks (penalty) and the final score all increase when the sense-cluster inventory is used (METEOR_SC). By using the French cluster information, METEOR manages to identify

[6] http://www.statmt.org/wmt09/evaluation-task.html

	METEOR	METEOR_SC
Matches	30477.92	**32320.25**
Chunks	17123.17	**18157.83**
Score	0.1250	**0.1326**

Table 2: *Statistics on the WMT09 EN–FR data (averaged over 12 systems).*

sense correspondences between the words found in the hypothesis and the reference which would otherwise be missed. In Table 3, we present some cases of matches captured by METEOR_SC.

4.3.2. Correlation with Human Judgments

In order to calculate the correlation that METEOR has with human judgments of translation quality during evaluation in French, we use the WMT09 evaluation shared task dataset which consists of translations of news stories. So, the test corpus is different from the corpus that was used for training the sense induction method (Europarl).

All English–French human rankings (390 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. We provide rankings at the segment and the system levels.

To measure the *correlation* of the automatic metrics with the human judgments of translation quality, we use Spearman's rank order correlation coefficient [27]. 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.

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

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.

In Table 4, we present the (sentence-level) results of the correlation of METEOR with human judgments, with and without the synonymy module, and we compare them to the results obtained for BLEU [22].

	Correlation
BLEU	0.3010±0.0481
METEOR	0.3477±0.0575
METEOR_SC	**0.3562±0.0562**

Table 4: *Average Segment Level Correlation results in French.*

According to these results, the correlation with human judgments increases when the sense-clusters are exploited by METEOR for evaluation in French (METEOR_SC) compared to BLEU and to METEOR with *exact* and *porter-stemmer* matches.

Furthermore, the system-level correlation results, presented in Table 5, show that METEOR is better than BLEU and METEOR_SC is better than METEOR.

	Correlation
BLEU	0.8462
METEOR	0.9021
METEOR_SC	**0.9161**

Table 5: *EN–FR System Level Correlation.*

4.4. Evaluation in English

4.4.1. MT evaluation

On the English side, we conducted experiments on the Metrics MATR 2008 development set (LDC2009T05) which consists of human-assigned *adequacy* scores to 1992 sentences generated by eight MT systems. The source data consists of twenty-five Arabic (AR) language newswire documents with a total of 249 segments.[7]

We perform exact, stemmed, synonymy and paraphrase matching using METEOR, and experiment with three types of synonymies: WordNet as in the original METEOR (WN), Sense Clusters generated by our algorithm (SC) and the combination of the previous two (WN+SC). We report the matching statistics in Table 6.

	WN	SC	WN+SC
Synonyms	267.13	356.25	498.78
Paraphrases	370.00	295.88	235.13
Score	0.7123	0.7140	0.7225

Table 6: *Statistics on the AR–EN data from LDC (averaged over 8 systems).*

According to these results, when using the sense-cluster inventory METEOR finds more matches than when WordNet is used, and the final score increases as well. A further increase in the final score is observed when the two semantic resources are combined, which shows that some of the matches that they identify are different.

We note that the paraphrases used by Meteor can bring a larger improvement on this data set. We suspect that this is partly due to the fact that our sense-clusters are derived from EN–FR bilingual data but METEOR's paraphrases are extracted from TERp paraphrases [28, 29], which are induced from AR–EN data (data also used to train many of the systems that we evaluate). We hypothesize that bilingually-motivated linguistic resources, such as our sense-clusters and

[7] The data in each segment includes four human reference translations in English and system translations from eight different MT systems.

Language	Hypothesis	Reference
FR	Tout comme le même prix de l'essence à la pompe des stations est **sûrement** le produit d'une **coïncidence** et de la difficile **lutte** concurrentielle.	De la même manière, un prix de l'essence strictement identique à toutes les stations de service est **surement** dû au **hasard** et au dur **combat** de la concurrence.
	Si le **voyageur** doit **confirmer** la validité des coupons enregistré, ...	Au cas oú le **passager** aurait besoin de **vérifier** la validité des coupons enregistrés, ...
	Ils devront **abandonner** leur relation actuelle avec l'Iran; ...	Ils vont devoir **renoncer** à leurs relations actuelles avec l'Iran; ...
	Mme Locklear a été arrêté soupçonné de conduite sous l'influence de certains - pas mieux substance spécifique, et **emprisonnés** dans la prison locale à 7.00pm, d'être **libérés** quelques heures plus tard.	Locklear a été arrêté étant suspectée d'avoir conduit sous l'effet d'une substance qui n'a pas été spécifiée et a été **incarcérée** vers 19.00 heures et **relâchée** quelques heures plus tard.
EN	According to Bush, the **project** should **address** the **main** causes of financial crisis and help stabilize the entire economy.	According to Bush, the **plan** would **tackle** the **basic** causes of the financial crisis and help stabilize the entire economy.
	The plan supports the financial system will be **examined** Monday in the House of Representatives.	The plan to support the financial system will be **discussed** in the House of Representatives on Monday.
	"We worked very hard on the issue and have made **significant** progress toward an agreement that will work and be **effective** for the market and to all Americans," said Paulson.	"We've worked very hard on this and we've made **great** progress toward an agreement that will work and that will be **useful** for all Americans," Paulson said.

Table 3: *Matches identified by METEOR when the sense-clusters are used (METEOR_SC).*

TERp paraphrases, work best on the language pairs they are trained on, and plan to validate and explore this hypothesis in future work.

4.4.2. Correlation with Human Judgments

The correlation results obtained on the Metrics MATR 2008 development set are given in Table 7. We experiment by using METEOR without the synonymy module (METEOR), with WordNet (WN), with sense-clusters (SC) and the combination of the last two (WN+SC), with and without paraphrases (+P/-P).

We calculate Pearson's correlation with human judgments on adequacy using Matthew Snover's correlation calculation tool.[8] Pearson's correlation is defined as in (2):

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

where x_i is the value of the i^{th} score, $\bar{X}$ is the mean score and s_X is the standard deviation.

As is shown by the Pearson's correlation and its confidence interval, our sense-clusters make generally the same level of contributions to Meteor as WordNet, no matter whether we apply a final round of paraphrase matching or not. With our sense-clusters, Meteor obtains a Pearson's correlation coefficient at the upper-bound of the 95% confidence interval of the version without it, which shows that the improvement introduced by our sense clusters is substantial. Here too, the highest correlation, with or without paraphrases, is obtained when WordNet and the sense-clusters are combined.

[8] `http://www.umiacs.umd.edu/~snover/terp/scoring/`

	Correlation	95% Interval
METEOR	0.7256	(0.704, 0.746)
WN-P	0.7455	(0.725, 0.764)
SC-P	0.7442	(0.724, 0.763)
WN+SC-P	0.7526	(0.733, 0.771)
WN+P	0.7634	(0.745, 0.781)
SC+P	0.7614	(0.742, 0.779)
WN+SC+P	0.7657	(0.747, 0.783)

Table 7: *Segment Level Correlation on the AR–EN data from LDC.*

The results of the AR–EN system-level correlation with human judgments are presented in Table 8. We observe that our SC marginally outperforms WN on system level, with or without paraphrases.

	Correlation	95% Interval
METEOR	0.96114	(0.795, 0.993)
WN-P	0.96647	(0.821, 0.994)
SC-P	0.96683	(0.823, 0.994)
WN+SC-P	0.96940	(0.835, 0.995)
WN+P	0.97630	(0.871, 0.996)
SC+P	0.97674	(0.873, 0.996)
WN+SC+P	0.97787	(0.879, 0.996)

Table 8: *AR–EN System Level Correlation.*

5. Analysis of the results

The results of the evaluation experiments demonstrate the positive impact of using the sense-cluster inventories during MT evaluation and, consequently, indicate the quality of these automatically generated resources. Exploiting the cluster inventory during evaluation in French renders the synonymy module of METEOR operable for evaluation in this language and permits to capture matches between the hypothesis and the references that would otherwise be missed. This is clearly shown by the difference in the evaluation scores obtained when the synonymy module is used and when it is omitted (cf. section 4.3). Furthermore, the use of the sense-clusters increases the correlation of the metric with human judgments of translation quality.

By using the English sense-cluster inventory, METEOR achieves the same correlation with human judgments of translation quality as with WordNet (cf. section 4.4). This similarity is important given that our inventory is automatically created while WordNet is hand-crafted. We should also note here the difference in the size and the coverage of the two resources. The English cluster inventory contains 11,775 clusters while WordNet counts 664,679 synsets. The results are even more encouraging if we consider the fact that the sense induction method is trained and tested on corpora of different domains : the training is performed on Europarl while the corpora used for testing contain news stories.

Another clear advantage of this automatically created semantic inventory, in comparison to a pre-defined resource, is that it can be further enriched and updated if more parallel corpora are available for training the sense induction method. As opposed to WordNet, which is often criticized as being too general for processing in specific domains, the sense-cluster inventory can easily be adapted to the domains of the processed texts. Actually, the similarity of the results obtained by using these two resources and their size difference could also serve to demonstrate the irrelevance of much of the information found in WordNet to specific domains of interest and, eventually, the redundancy characterizing this resource.

By revealing the benefit that can be achieved during MT evaluation when the sense-cluster inventories are used, this extrinsic evaluation clearly demonstrates the quality of their contents and their usefulness in this application setting.

6. Conclusion

In this paper, we have presented a clustering algorithm for cross-lingual unsupervised sense induction from parallel corpora. We have analyzed the properties of the algorithm and described the bilingual sense-cluster inventories that were generated from a parallel training corpus. Given the problems posed by an intrinsic evaluation of the obtained semantic resources, an extrinsic large-scale quantitative evaluation was carried out by exploiting them in a MT evaluation setting. The evaluation results prove that the sense-cluster information present in the resources permits to efficiently capture cases of lexical variation during MT evaluation, to an extent comparable to that observed when a large hand-crafted resource is used.

As part of future work, we intend to generate sense-cluster inventories from different data sets. Given that the sense induction method is language-independent, cluster inventories can be created in other languages not disposing of large semantic resources for rendering MT evaluation more sensitive to semantics and lexical variation. We would also like to compare our results to those that would be obtained by using another WordNet-like resource for French [30], currently under development. Moreover, we plan to integrate the inventories in other multilingual applications. The sense-cluster inventories can be efficiently used for cross-lingual WSD. So, an important avenue for future work is to integrate a WSD method exploiting our resources into a Statistical MT system.

7. Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.

8. References

[1] R. Mihalcea and D. I. Moldovan, "Automatic generation of a coarse grained WordNet," in *Proceedings of the 14th FLAIRS Conference*, 2001, pp. 454–458.

[2] P. Edmonds and A. Kilgarriff, "Introduction to the special issue on evaluating word sense disambiguation systems," *Natural Language Engineering*, vol. 8, no. 4, pp. 279–291, 2002.

[3] N. Ide and Y. Wilks, "Making Sense About Sense," in *E. Agirre and P. Edmonds (eds.), Word Sense Disambiguation, Algorithms and Applications*, Springer, 2007, pp. 47–73.

[4] P. Resnik and D. Yarowsky, "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation," *Natural Language Engineering*, vol. 5, no. 3, pp. 113–133, 2000.

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

[6] G. Miller and C. Fellbaum, "WordNet," 2007, http://wordnet.princeton.edu/.

[7] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments," in *Proceedings of the*

ACL-2007 Workshop on Statistical Machine Translation, Prague, Czech Republic, 2007, pp. 228–231.

[8] M. Apidianaki, Y. He, and A. Way, "Capturing lexical variation in MT evaluation using automatically built sense-cluster inventories," in *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*, Hong Kong, 2009, pp. 53–62.

[9] H. Schütze, "Automatic Word Sense Discrimination," *Computational Linguistics*, vol. 24, no. 1, pp. 97–123, 1998.

[10] T. Pedersen and R. Bruce, "Distinguishing Word Senses in Untagged Text," in *Proceedings of the Second EMNLP Conference*, Providence, R.I., 1997, pp. 197–207.

[11] P. Pantel and D. Lin, "Discovering word senses from text," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, 2002, pp. 613–619.

[12] C. Cabezas and P. Resnik, "Using WSD Techniques for Lexical Selection in Statistical Machine Translation," in *Technical report CS-TR-4736/LAMP-TR-124/UMIACS-TR-2005-42*, 2005.

[13] M. Carpuat and D. Wu, "Improving Statistical Machine Translation using Word Sense Disambiguation," in *Proceedings of the Joint EMNLP-CoNLL Conference*, Prague, Czech Republic, 2007, pp. 61–72.

[14] P. Jin, Y. Wu, and S. Yu, "SemEval-2007 Task 5: Multilingual Chinese-English Lexical Sample," in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, 2007, pp. 19–23.

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

[16] N. Ide, T. Erjavec, and D. Tufis, "Sense discrimination with parallel corpora," in *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, 2002, pp. 54–60.

[17] M. Apidianaki, "Translation-oriented sense induction based on parallel corpora," in *Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco, 2008, pp. 3269–3275.

[18] G. Grefenstette, *Explorations in Automatic Thesaurus Discovery*. Dordrecht : Kluwer Academic Publisher, 1994.

[19] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Proceedings of MT Summit X*, Phuket, Thailand, 2005, pp. 79–86.

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

[21] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

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

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

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

[25] L. Zhou, C.-Y. Lin, and E. Hovy., "Re-evaluating Machine Translation Results with Paraphrase Support," in *Proceedings of EMNLP*, Sydney, Australia, 2006, pp. 77–84.

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

[27] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "Further Meta-Evaluation of Machine Translation," in *Proceedings of the ACL-2008 Workshop on Statistical Machine Translation*, Columbus, Ohio, 2008, pp. 70–106.

[28] C. Callison-Burch, "Syntactic Constraints on Paraphrases Extracted from Parallel Corpora," in *Proceedings of EMNLP*, Honolulu, Hawaii, 2008, pp. 196–205.

[29] M. Snover, N. Madnani, B. Dorr, and R. Schwartz, "TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate." *Machine Translation*, vol. 23, no. 2-3, pp. 117–127, 2009.

[30] B. Sagot and D. Fišer, "Building a free French wordnet from multilingual resources," in *Proceedings of Ontolex*, Marrakech, Morocco, 2008.