Measuring the adequacy of cross-lingual paraphrases in a Machine Translation setting

Marianna APIDIANAKI
LIMSI-CNRS
BP 133, 91403 Orsay Cedex
France
marianna@limsi.fr

ABSTRACT
Following the growing trend in the semantics community towards models adapted to specific applications, the SemEval-2 Cross-Lingual Lexical Substitution and Word Sense Disambiguation tasks address the disambiguation needs of Machine Translation (MT). The experiments conducted in this study aim at assessing whether the proposed evaluation protocol and methodology provide a fair estimate of the adequacy of cross-lingual predictions in translations. For this purpose, the gold SemEval paraphrases are fed into a state-of-the-art MT system and the obtained translations are compared to paraphrase quality judgments based on the source context. The results show the strong dependence of cross-lingual paraphrase adequacy on the translation context and cast doubt on the contribution that systems performing well in existing evaluation schemes would have on MT. These empirical findings highlight the importance of complementing the current evaluation schemes with translation information to allow a more accurate estimation of the systems impact on end-to-end applications.

KEYWORDS: Cross-Lingual Word Sense Disambiguation, Cross-Lingual Lexical Substitution, paraphrasing, Machine Translation.

1 Introduction

An important trend in computational semantics in recent years is the adaptation of inventories, models and evaluations to specific applications. In this vein, the Cross-Lingual Lexical Substitution (CLLS) and Word Sense Disambiguation (CL-WSD) tasks of SemEval-2 address the disambiguation needs of multilingual applications: what is being evaluated is the capacity of the participating systems to provide semantically correct translations for words in context that could, among others, constitute the input of Machine Translation (MT) systems (Mihalcea et al., 2010; Lefever and Hoste, 2010).[1] The underlying assumption is that the closer the output of a CLLS/CL-WSD system is to a manually built gold standard of cross-lingual paraphrases, the higher its contribution in a real application will be.

Paraphrasing is highly useful in MT as is shown by the substantial amount of research undertaken on the subject.[2] It permits to deal with out-of-vocabulary words (Callison-Burch et al., 2006; Marton et al., 2009), capture lexical variation during evaluation (Zhou et al., 2006; Owczarzak et al., 2006), expand the set of reference translations for minimum error rate training (Madnani et al., 2007) and improve the general performance of MT systems (Max, 2010). It is however interesting that in spite of the MT orientation of the CL SemEval-2 tasks, translation selection and evaluation are carried out by reference solely to the source language. The target language context which plays an important role in lexical selection in statistical MT systems, as highlighted by the strong influence of the language model on word choice, is not considered.

In this work, we explore the role of the target language in CLLS and CL-WSD by measuring the adequacy of CL paraphrases in translations. Our goal is not to estimate the impact of semantics in MT, as was the case in previous works on the subject (Carpuat and Wu, 2007; Chan et al., 2007), but to empirically test the adequacy of the sense descriptions provided in the CL evaluation tasks in an MT setting. The paper is organized as follows. The CL SemEval-2 tasks are described in Section 2. The adopted experimental methodology and evaluation setup are presented in Section 3. The analysis of the obtained results, in Section 4, highlights the importance of the target language context for CLLS and CL-WSD, and the implications of these findings for CL semantic evaluations.

2 Translation context in cross-lingual semantic evaluations

2.1 The SemEval-2 Cross-Lingual tasks

In the CLLS and CL-WSD tasks of the SemEval-2 evaluation campaign, the participating systems had to predict semantically correct translations in different languages for English target words in context (Mihalcea et al., 2010; Lefever and Hoste, 2010). The performance of the systems was measured by comparing their output to a manually built gold standard (GS) of cross-lingual paraphrases. For example, the instance of the target word *fresh* in sentence #952 of the CLLS test set: "*At first the user is impressed by the **fresh** clean smell coming out of the machine and how nice it makes their home smell.*", was tagged by the following set of translations which express the sense of *fresh* in Spanish: *fresco 4; puro 1; flamante 1; limpio 1; nuevo 1*. GS translations are lemmatized and the frequency counts indicate the number of annotators that proposed each substitute.

The differences between the two tasks mainly lie in the targeted lexical samples and the involved

[1] These systems can also help human translators in their work and assist language learners.

[2] See (Madnani and Dorr, 2010) for a comprehensive survey of data-driven methods for paraphrase generation.

language pairs. CLLS addresses words of all open-class parts of speech in one language pair (English-Spanish) while CL-WSD focuses on the translation of English nouns in five languages (French, Spanish, German, Dutch and Italian).[3] Another point of variation concerns the definition of senses. In CL-WSD, target word senses were described by means of clusters of their semantically similar translations (Ide et al., 2002; Apidianaki, 2008). More precisely, the translations of the target words in the Europarl corpus (Koehn, 2005) were manually clustered and the obtained clusters served for tagging. On the contrary, CLLS did not involve a clustering step and the annotators could propose translations found in any external resource. The CLLS test set was built from the English Internet Corpus (Sharoff, 2005) while CL-WSD test sentences were extracted from the BNC[4] and the JRC-ACQUIS corpus (Steinberger et al., 2006).

2.2 Translation context: a neglected parameter

Although the CL SemEval tasks are clearly oriented towards MT, annotator judgments and system suggestions are made on the basis of source language information. Translations are selected so as to express the meaning of the target words in the target language but the translation context in which they would be used has no influence on the selection process. This lack of target language information would have a minimal impact in settings where CLLS/CL-WSD systems serve to assist human users, but becomes more important in the context of MT where the proposed CL paraphrases have to be automatically filtered to select the most adequate translation. This selection is not straightforward for several reasons.

Words that seem interchangeable on the basis of formal criteria, such as distributional similarity, might not be substitutable in real texts because of other parameters preventing the substitution (e.g. syntactic structure, collocations). In a translation setting where the substitution is done cross-lingually, it is important that the paraphrases preserve both the sense of the original word (or phrase) and the fluency of the translated text. However, clustered translations are usually near-synonyms translating the same sense, but almost never absolute synonyms interchangeable in translations (Edmonds and Hirst, 2002; Apidianaki, 2009). Consequently, although CLLS and CL-WSD could greatly contribute in MT by enhancing the semantic relevance of translations, the existing evaluations do not provide a fair estimate of the systems' capacity to propose translations that would also fit well in the translated texts.

We conduct a series of experiments to assess the adequacy of CL paraphrases in translations by exploiting the CLLS and CL-WSD test sets. As the two test sets were mainly built from monolingual corpora, no reference translations are available against which the quality of the CL paraphrases could be measured using standard MT evaluation metrics (BLEU, METEOR, etc.). So, we adopt a variation of the *substitution-based* approach used in works on paraphrasing (Bannard and Callison-Burch, 2005) for validating candidate paraphrases, based on the assumption that items deemed to be paraphrases may behave as such only in some contexts and not in others. We translate the CLLS and CL-WSD test sets with a state-of-the-art MT system by exploiting the manually-defined GS paraphrases. Once the set of translations for each test sentence is produced, we measure the substitutability of the GS paraphrases using an automatic and a human ranking, as explained in the next section.

[3] The CLLS lexical sample is composed of 300 noun, 310 verb, 280 adjective and 110 adverb instances with approximately 5 Spanish substitutes per target word and a pairwise inter-annotator agreement of 0.2777. The CL-WSD test data contains 50 instances of 20 target nouns and their substitutes in five languages.

[4] http://www.natcorp.ox.ac.uk/

3 Experimental setup

3.1 Systems and data

The CLLS and CL-WSD test sets are translated into Spanish and French, respectively, using the baseline system of the WMT-2011 shared task (Moses) (Koehn et al., 2007). The two MT systems are trained on the data released for WMT-2011 for the two language pairs, namely the French-English and Spanish-English parts of Europarl (version 6) (Koehn, 2005). The language models used during decoding are trained on the monolingual Spanish and French parts of Europarl. For each test sentence, we constrain the decoder to produce translations by using all GS paraphrases. These are plugged into Moses using its 'XML Markup' feature which allows to specify translations for parts of the input sentence. The 'exclusive' mode is activated which forces the decoder to use the XML-specified translations and ignore any phrases from the phrase table that overlap with that span.[5] In total, 4,791 unique Spanish translations are produced for the CLLS test set and 4,220 French translations for the CL-WSD test set.

The GS paraphrases are lemmatized, so we first produce translations at the lemma level without dealing with inflections. At this stage, the test sentences are lemmatized and the MT systems are trained on lemmatized bi-texts. The CL-WSD test set is also translated into French using inflections. We gather all the inflectional variants of each paraphrase found in the training bi-text and provide them to Moses through the XML markup. For instance, to translate the test sentence: "*Taking with determination this road leading to a dynamic European Union on the world* ***scene*** *will yield further substantial benefits to all parties involved in the EU and beyond.*" we provide all inflected forms of each GS paraphrase found in Europarl: *scène/scènes, niveau/niveaux, marché/marchés*, etc. The MT system then selects the best inflection depending on the surrounding context, as shown in Table 1.

3.2 Automatic ranking

The set of lemmatized translations produced by Moses for each test sentence is ranked by a target language model (lm). Language model scores reflect the probability of the sentences formed by substituting paraphrases and are useful for ranking candidate paraphrases in automatic paraphrasing tasks. Bannard and Callison-Burch (2005), for example, combine a language model probability with a paraphrase probability to rank candidate paraphrases produced by the *pivot* method.[6] The use of a language model allows to account for the fact that the best paraphrase might vary depending on information about the sentence it appears in and lets the surrounding words in the sentence influence paraphrase ranking and selection.

We build two extended lms (in Spanish and French) using additional monolingual data compared to that used for training the lms used by Moses. The training data comprises Europarl, the News Commentary corpus and the 2009, 2010 and 2011 News Crawl data provided at the WMT-11 shared task for the two languages. We employ the SRILM toolkit (Stolcke, 2002) to compute two 5-gram language models and, subsequently, to score and rank the translations produced by Moses. As the use of different GS paraphrases may alter the context of the translated sentences normalized lm scores are used, defined as $\frac{1}{n} - \log(P)$, where n is the length of the translation

[5] The 'inclusive' mode allows phrase table entries to compete with the XML entry. This configuration permits to define probabilities for the provided translation choices and leave the final selection to the target language model.

[6] In the pivot method, phrases in one language are considered to be potential paraphrases of each other if they share a translation in another language. The paraphrase probability is defined in terms of the translation model probabilities that the original phrase translates as a particular phrase in the other language.

GS	Translation	lm score
scène (3)	prendre avec détermination cette voie conduisant à une dynamique de l' union européenne sur la **scène** mondiale enregistre encore des avantages substantiels pour toutes les parties concernées dans l' ue et au-delà .	2
niveau (2)	... menant à une union européenne dynamique au **niveau** mondial engendrera davantage des avantages substantiels pour toutes les parties concernées ...	2.13
vie (2)	... conduisant à une dynamique de l' union européenne sur la **vie** apportera des avantages substantiels pour toutes les parties impliquées ...	1.8
marché (1)	... conduisant à une dynamique de l' union européenne sur le **marché** mondial engendrera davantage des avantages substantiels pour toutes les parties concernées ...	1.96
plan (1)	... menant à une union européenne dynamique sur le **plan** mondial engendrera davantage des avantages substantiels pour toutes les parties concernées ...	2.09

Table 1: Ranking of Moses translations using GS paraphrases and lm scores.

and *P* the language model probability. Table 1 shows the normalized lm scores of the set of translations produced for the test sentence given in the previous section.[7] The lm ranking is compared to the GS one which reflects the semantic relevance of the paraphrases as estimated by reference to the source context. Our hypothesis is that a high correlation between the two rankings would indicate that translations privileged in the GS (i.e. with a high frequency) would serve to produce fluent translations (i.e. with better lm scores). Given the important role of lms in lexical selection, the low ranking of paraphrases could be interpreted as denoting their lower chances of being used in translations. However, this judgment cannot be absolute as the language model is one among other components that determine lexical choice in MT systems.

3.3 Human ranking

Although the lm scoring yields interesting results, we consider that it is not reliable enough to lead to safe conclusions as to the adequacy of CL paraphrases in translations. So, we also conduct a human evaluation. The annotators are asked to rank the set of Moses translations produced for each target word instance on a 3-point scale, according to the adequacy of the paraphrases and the fluency of the translated text.[8] Good quality paraphrases (i.e. the highest ranked ones, assigned a '1' value) should preserve both the meaning of the source word and the grammaticality of the target sentence. This experiment can be viewed as a substitution test (Callison-Burch, 2008) with the difference that the paraphrases are not just substituted in the translated sentences but fed into the MT system which exploits them during translation. Consequently, the context surrounding the paraphrase might be altered as well, as shown in the examples given in Table 1.

The human ranking covers 538 instances of the CL-WSD test set with an average of 4.17 French paraphrases per instance. The 538 translation sets produced by Moses contain a total of 1821

[7] Normalized scores are rounded to two decimal places. Translations with lower scores are considered as more fluent.

[8] The annotators are native and highly proficient French speakers working on MT and paraphrasing.

unique translations and each translation is annotated twice. We calculate the inter-annotator agreement using Cohen's kappa coefficient for three different annotation configurations: the ranking performed using the 3-point scale and two coarser-grained rankings obtained by interpreting intermediate ('2') values as denoting good or low quality translations (i.e. converting them into '1's or '3's). As shown by the kappa values given in Table 2, agreement on the 3-point ranking is rather low ($K = 0.35$) but it gets higher when the intermediate values are interpreted as 'good' or 'bad'. In the first case kappa is 0.57, which is considered as substantial agreement, but it reaches its highest value ($K = 0.72$) when medium-ranked translations are considered as low quality ones (2→3). This practically means that in most cases both annotators perceive a problem in the translated texts but have a different estimate of its severity. The increase of the kappa value when a scale with fewer points is used is natural and has been observed in other works on paraphrasing.[9]

rating scale	kappa
3-point scale	0.35
2-point scale (2 → 1)	0.57
2-point scale (2 → 3)	**0.72**

Table 2: Inter-annotator agreement.

Examples of human-ranked translation sets are given in Table 3. We observe that medium-ranked CL paraphrases, such as the translation ***charge*** of the target word *strain* (assigned values '2' and '3') or the translation ***parties*** of the noun *side*, do not fit well in the translated texts. However, the annotators give some credit to paraphrases that may seem awkward in the translated texts but still carry some of the semantic load of the source word, reserving the lowest values to erroneous translations from both points of view. Given the inadequacy of medium-ranked paraphrases in translations, we consider these judgments as low quality ones and distinguish between two categories. The $K = 0.72$ agreement obtained in this case is very high, especially for a semantics task like this one.

4 Results

4.1 Gold standard judgments *vs* language model scores

We calculate the correlation of the two rankings with the GS frequency ranking. We first compute the correlation between the semantic relevance of CL paraphrases, as reflected in the GS frequencies, and their adequacy in translated texts, as measured by the lm. We use the Spearman's rank order correlation coefficient (ρ), a non-parametric test, because the data does not seem to be normally distributed. The Spearman coefficient is defined as the Pearson correlation between ranked variables. To compute the correlation of two random variables X and Y, Pearson's coefficient divides their covariance by the product of their standard deviations.

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

To compute Spearman's ρ, absolute values are transformed into ranks.[10] The correlation

[9] Callison-Burch (2008) reports a kappa agreement of 0.33 when a 5-point scale is used and an agreement of 0.61 with a 2-point scale. The scale conversion is performed by measuring agreement in terms of how often the annotators assigned a value higher or lower than a pre-defined threshold.

Word	Source	Translations	Ranks
strain	Exposure both in working life and everyday living to different sets of values, assumptions, expectations, and behaviour patterns places a severe **strain** on the individual.	l' exposition à la fois dans la vie professionnelle et de la vie de tous les jours à différents ensembles de valeurs , des hypothèses , de leurs attentes et leurs comportements accorde une très forte **pression** sur les individus	1|1
		... lieux de graves **tensions** sur les individus	2|2
		... peser une **charge** sur les individus	2|3
		... peser une grave **pesant** sur les individus	3|3
		... peser une grave **serrée** sur l' individu	3|3
		... peser une grave **grevée** sur l' individu	3|3
side	Many American students working in British drama schools find the answer to this question by using what is called "standard American", and this approach is being used now in training on both **sides** of the Atlantic.	bon nombre des étudiants américains travaillent dans les écoles du drame , trouver la réponse à cette question , en utilisant ce qui est appelé " norme américaine " , et cette approche est utilisé dans la formation sur les deux **rives** de l' atlantique	1|1
		... des deux **côtés** de l' atlantique	1|1
		... des deux **bords** de l' atlantique	1|3
		... des deux **parties** de l ' atlantique	2|3
		... des deux **transatlantique** de l' atlantique	2|3
		... des deux **outre** de l' atlantique	3|3

Table 3: Manually ranked translations.

between the GS annotations in the French data set and the lm scores on the lemmatized translation dataset is $\rho = 0.067$ and highly significant with $p = 1.361e{-}05$ (< 0.05). Spearman correlation with the normalized lm scores is $-.014$ with a p-value of .363. As the dataset with the normalized scores contains *ties*, we also calculate the Kendall's tau-b non-parametric correlation. Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be a set of joint observations from two random variables X and Y, the Kendall's tau coefficient is defined as

$$\tau = \frac{(|\text{concordant pairs}| - |\text{discordant pairs}|)}{\frac{1}{2}n(n-1)} \qquad (2)$$

Concordant is any pair of observations (x_i, y_i) (x_j, y_j) where the ranks for both elements agree (e.g. $x_i > x_j$ and $y_i > y_j$), otherwise it is discordant. Kendall's tau-a requires all the values of x_i and y_i to be unique for the p-value to be accurate, but Kendall's tau-b accounts for ties (i.e. pairs of observations where $x_i = x_j$ or $y_i = y_j$) .[11] The Kendall's tau-b correlation between the GS ranking and the normalized lm scores is low: $-.011$ ($p = .363$). This lack of correlation could mean in practical terms that the best paraphrases from a semantics point of view would not lead to more fluent translations. To draw safer conclusions we present in the next section the results obtained by the human ranking.

The correlation between GS estimates and the unnormalized lm scores for Spanish is $\rho = 0.0242$, with lower significance than in French ($p = 0.09$). Given the similar size of the test sets in the two languages, this divergence might be due to the higher homogeneity of the French dataset

[10] The analysis is done using the R package: `http://www.r-project.org`

[11] Kendall's tau-b correlation is calculated using the IBM SPSS statistics environment.

which contains only nouns. Words of different parts of speech, found in the Spanish test set, are handled differently by the annotators and their paraphrases have a varying impact on translation fluency. The correlation computed between the Spanish GS scores and the normalized lm scores is low as well, with $\rho = .005$ ($p = .726$) and a Kendall's tau-b value of .004 ($p = .723$).

4.2 Gold standard *vs* target language human judgments

The dataset that consists of the GS frequency estimates and the human judgments of translation adequacy contains ties, so we calculate the Kendall's tau-b correlation. We use the values assigned in the first annotation pass. The obtained correlation is $-.271$, for the 3-point scale (negative because the values in the two rankings are inverted), and $-.26$ for the 2-point scale (conversion 2→3). Both correlations are significant at the 0.01 level. The 3-point scale judgments correlate slightly better with the GS ones because they are rated on the same scale.[12] These results show that paraphrases privileged in the GS do not fit well in the translated texts, while translations ranked low in the GS might be preferred in translations.

We finally calculate the correlation between the human ranking and the normalized lm scores on unlemmatized translations. Kendall's tau-b correlation is .018 and .033, for the 3 and the 2-point scale respectively, but the p-values are quite high (.334 and .091). It would be interesting to repeat this correlation experiment once more annotated examples will be available. A detailed analysis of this discordance would provide valuable hints on the capacity of lms to measure fluency and paraphrase adequacy. We observe, for instance, that the annotators often base their judgments on the context surrounding the paraphrases although lm scores are computed on the entire sentences that might be altered during translation. Nevertheless, the fact that these correlation results are not yet safe does not influence the conclusions that can be drawn from the low correlation observed between the gold standard ranking and the human ranking of translation adequacy, which is highly significant.

Conclusion

The findings of this study reveal that the results of the CL SemEval-2 tasks are not indicative of the contribution that the participating systems would have in MT. It has been shown that although the proposed evaluation metrics address the semantic relevance of CL paraphrases, they do not account for their suitability in translations. These empirical results highlight the importance of integrating translation information in CL semantic evaluations by resorting either to simplified translation tasks (Vickrey et al., 2005) or to full-fledged MT systems. Evaluation metrics capable of rewarding semantically correct translations that do not distort the fluency of the translations are much needed in the field of MT for evaluating the output of MT systems and the contribution of disambiguation modules. Another perspective worth exploring is the set up of all-words CL evaluation tasks, in addition to the lexical sample ones, allowing to assess the global capacities of CLLS and CL-WSD systems and the coverage they can attain in real-life applications. This setting would also permit to explore the potential of collaboration between CL-WSD modules and MT systems for correct lexical selection.

Acknowledgments

We would like to thank the annotators for the time and effort they put into this task. This work was partly funded by the Cap Digital SAMAR project.

[12] GS paraphrases were assigned frequencies from '1' to '3'; only two cases were assigned a '4' value.

References

Apidianaki, M. (2008). Translation-oriented sense induction based on parallel corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-08)*, pages 3269–3275, Marrakech, Morocco.

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

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics.

Callison-Burch, C. (2008). Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 196–205, Honolulu, Hawaii. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA. Association for Computational Linguistics.

Carpuat, M. and Wu, D. (2007). Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of the Joint EMNLP-CoNLL Conference*, pages 61–72, Prague, Czech Republic.

Chan, Y. S., Ng, H. T., and Chiang, D. (2007). Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic. Association for Computational Linguistics.

Edmonds, P. and Hirst, G. (2002). Near-synonymy and lexical choice. *Computational Linguistics*, 28:105–144.

Ide, N., Erjavec, T., and Tufiş, D. (2002). Sense discrimination with parallel corpora. In *Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual ACL Meeting, Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Lefever, E. and Hoste, V. (2010). Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden. Association for Computational Linguistics.

Madnani, N., Ayan, N. F., Resnik, P., and Dorr, B. J. (2007). Using Paraphrases for Parameter Tuning in Statistical Machine Translation. In *Proceedings of the Workshop on Statistical Machine Translation*, Prague, Czech Republic. Association for Computational Linguistics.

Madnani, N. and Dorr, B. J. (2010). Generating Phrasal and Sentential Paraphrases: A Survey of Data-driven Methods. *Computational Linguistics*, 36:341–387.

Marton, Y., Callison-Burch, C., and Resnik, P. (2009). Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Singapore.

Max, A. (2010). Example-based paraphrasing for improved phrase-based statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 656–666, Cambridge, MA. Association for Computational Linguistics.

Mihalcea, R., Sinha, R., and McCarthy, D. (2010). Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden. Association for Computational Linguistics.

Owczarzak, K., Groves, D., Van Genabith, J., and Way, A. (2006). Contextual bitext-derived paraphrases in automatic mt evaluation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 86–93, New York City. Association for Computational Linguistics.

Sharoff, S. (2005). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11:435–462.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., and Tufiş, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147, Genoa, Italy.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-02)*, pages 901–904, Denver, CO.

Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (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)*, pages 771–778, Vancouver, Canada.

Zhou, L., Lin, C.-Y., and Hovy, E. (2006). Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 77–84, Sydney, Australia. Association for Computational Linguistics.