Alignment-based sense selection in METEOR and the RATATOUILLE recipe

Benjamin Marie
LIMSI-CNRS, Orsay, France
Lingua et Machina, Le Chesnay, France
benjamin.marie@limsi.fr

Marianna Apidianaki
LIMSI-CNRS, Orsay, France
marianna@limsi.fr

Abstract

This paper describes Meteor-WSD and RATATOUILLE, the LIMSI submissions to the WMT15 metrics shared task. Meteor-WSD extends synonym mapping to languages other than English based on alignments and gives credit to semantically adequate translations in context. We show that context-sensitive synonym selection increases the correlation of the Meteor metric with human judgments of translation quality on the WMT14 data. RATATOUILLE combines Meteor-WSD with nine other metrics for evaluation and outperforms the best metric (BEER) involved in its computation.

1 Introduction

The Meteor metric evaluates translation hypotheses by aligning them to reference translations and calculating sentence-level similarity scores (Banerjee and Lavie, 2005; Denkowski and Lavie, 2010). The space of possible alignments for a hypothesis-reference pair is constructed by identifying all possible matches between the sentences according to different matchers mapping words with identical surface forms or having the same stem, WordNet synonyms and paraphrases. These modules add flexibility to the metric and improve its correlation with human judgments of translation quality but they fail to account for important semantics-related aspects. For example, Meteor and Meteor-NEXT treat all the variants available for a particular text fragment in WordNet (Fellbaum, 1998) or a pivot paraphrase database (Bannard and Callison-Burch, 2005) as semantically equivalent. Consequently, erroneous matches can be made by mapping synonyms found in different WordNet synsets and describing different senses. Similarly, pivot paraphrase sets merge sense boundaries in cases of polysemous words (Apidianaki et al., 2014), which means that paraphrases of different senses are considered as equivalent and can be mapped during evaluation. To avoid erroneous matches between text segments, it is thus important to restrict the available word and phrase variants to the ones that are correct in a specific context.

Context-based synonym selection is the main idea behind the Meteor-WSD metric submitted to the WMT15 Metrics Shared Task. The mechanism used for sense selection is described in detail in the next section where we also present the results obtained by the Meteor-WSD metric on the WMT14 evaluation dataset. Section 3 presents the RATATOUILLE metric which integrates Meteor-WSD together with nine other evaluation metrics. We report results in all language pairs and directions of the WMT14 dataset, except for hi-en.

2 Meteor-WSD

2.1 Context-dependent sense selection

A first attempt to integrate context-based sense selection in Meteor is described in Apidianaki and Marie (2015). Word sense disambiguation (WSD) was performed using the Babelfy tool (Moro et al., 2014) which relies on the multilingual resource BabelNet (Navigli and Ponzetto, 2012). BabelNet is a wide coverage semantic network where senses are described by synsets (synonym and paraphrase sets) containing lexicographic and encyclopedic knowledge extracted from various sources in many languages and are linked between them by different types of relations. Depending on the language, the lexical and phrase variants available in the synsets come from different sources such as WordNet, Wikipedia, Wiktionary, OmegaWiki as well as Machine Translation output. The Babelfy tool jointly performs WSD and Entity Linking by exploiting BabelNet's graph structure and se-

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 385–391,
Lisboa, Portugal, 17-18 September 2015.

lects multilingual BabelNet synsets that correctly describe the semantics of words in context.[1] In Apidianaki and Marie (2015), Babelfy assigned BabelNet synsets to words in the English references of the WMT14 dataset. The WordNet literals found in the synset selected for an English word served to filter the WordNet synonym set used by the basic Meteor configuration in order to keep only variants that were good in this specific context and discard the ones corresponding to other senses. The reported MT evaluation results showed the beneficial impact of disambiguation which improved the correlation of the metric to human judgments from almost all languages involved in the WMT14 evaluation into English (except for Czech-English). Naturally, performance strongly depends on the quality of the WSD annotations.

In this work, we use a recent version of the alignment-based WSD method proposed by Apidianaki and Gong (2015) which gives better disambiguation results than Babelfy on the WMT14 data. Disambiguation is now applied to references of all languages in the data, not only in English. The WSD method used in our experiments still relies on alignments but implements a mechanism that improves WSD in languages other than English compared to the previous version. More precisely, Apidianaki and Gong (2015) showed that the problematic sorting performed by the default BabelNet sense ranking mechanism in languages other than English has a strong negative impact on WSD. [2] In our experiments, we implement an alternative solution that eliminates the need for sense ranking. Furthermore, the currently used version integrates a multiword expression (MWE) identification step prior to disambiguation.

2.2 Data preparation

The WMT14 shared task involved five language pairs: English-French / German / Czech / Russian / Hindi. We provide results for all languages except for Hindi, and for both translation directions. Source and reference texts are lemmatised and part-of-speech tagged using the TreeTagger (Schmid, 1994), except for Czech where the MorphoDiTa tool (Straková et al., 2014) is used. The texts are then aligned at the lemma level using GIZA++ (Och and Ney, 2003).

2.3 Alignment-based MWE extraction

We identify candidate multiword expressions in the reference texts prior to disambiguation using word alignments and filter them using information in the BabelNet resource (version 2.5).[3] We consider as a candidate MWE a sequence of words in one language that is aligned to a single word in the other language (a $n:1$ alignment).[4] For example, *téléphone portable* is considered as a candidate French MWE because both its parts are aligned to *cellphone*. We validate a candidate MWE if it constitutes a separate entry in the BabelNet resource either in its lemmatised or in its unlemmatised form (retrieved from the text), otherwise we discard it. This procedure eliminates many noisy MWEs but some good ones are also left out because they are not present in the resource.

If a BabelNet entry is found for the MWE, the variants provided in the corresponding synset are extracted. For instance, we extract *téléphone mobile*, *téléphone cellulaire*, and *GSM* as variants of *téléphone portable*. The variants retrieved from BabelNet are used to annotate the instances of the MWEs in the reference texts. A validated MWE is thus considered as a unit and is excluded from disambiguation. The WSD step, that follows, assigns a sense to all content words (nouns, verbs, adjectives and adverbs) in the reference text that were not identified as part of a MWE.

2.4 Alignment-based disambiguation

The procedure for selecting the most adequate BabelNet synset for an occurrence of a word (w) in context is described in Figure 1. First, we find the synsets of w (S_w) in BabelNet 2.5 and filter them to keep only synsets that contain both w and its aligned translation t in this context ($S_w^t \subseteq S_w$). If only one synset is retained, we keep the variants (synonyms and paraphrases) of the same language as w provided in this synset. If several

[1] The Babelfy API can be downloaded from `http://babelfy.org`

[2] The BabelNet API sorts English senses according to their frequencies in WordNet, which are calculated from the sense annotated English corpus SemCor. As frequency information is not available for languages other than English, the BabelNet API sorts senses in lexicographic order, a criterion that fails to reflect their importance.

[3] The resource can be found at `http://babelnet.org` together with detailed statistics regarding the number of lemmas, senses and named entities provided, and the knowledge sources that were exploited for each language. Note that BabelNet's coverage varies a lot across languages.

[4] In future work, we intend to extend this heuristic to $n:m$ alignments linking sequences of two or more words in the two languages as in de Caseli et al. (2010).

Notation:
S_w: the set of BabelSynsets for w
t: a translation of w in context
S_w^t: the set of synsets in which t appears
V_w: the set of synonyms/paraphrases of w
l: language

The Sense Selection Algorithm:
$S_w^t \leftarrow \emptyset$
$S_w \leftarrow$ `getBabelSynsets`(w)
for each BabelSynset $s \in S_w$ **do**
  if $t \in s$ **then**
    add s to S_w^t
if $|S_w^t| \geq 1$ **then**
  for each BabelSynset $s \in S_w^t$ **do**
    $V_w \leftarrow$ `getVariants`(s, l)
  return (V_w)
else
  if l = English **then**
    $V_w \leftarrow$ `getVariants`(`getBFS`(S_w, l) , l)
  else
    for each BabelSynset $s \in S_w$ **do**
      $V_w \leftarrow$ `getVariants`(s, l)
    return (V_w)

Figure 1: The `getBabelSynsets` function retrieves the synsets available for w in BabelNet. The `getVariants` function returns the variants of w in the same language found in the synsets. If no synset is retained through alignment, the system falls back to the BFS baseline. The `getBFS` function ranks English synsets according to importance and returns the most frequent one (BabelNet First Sense).

synsets are retained, we keep the variants found in all synsets. Given the fine granularity of BabelNet senses (similar to WordNet), the intuition behind this merge is that different synsets containing the word and its translation describe closely-related senses.[5] Grouping the synsets that contain the aligned translation eliminates the need for sense sorting which is problematic in languages other than English, as explained in Section 2.1.

The system falls back to the most frequent sense provided by the default sense comparator of the BabelNet 2.5 API (`BabelSynsetComparator`) for unaligned English words or when the aligned translation is not found in any synset. To avoid applying the sense sorting procedure to languages other than English, we keep all available synsets for unaligned words in these languages or for words whose alignment is not found in any synset. In these cases, variants from all synsets are grouped together and no disambiguation is performed.

Disambiguation is applied to all content words in the texts (nouns, verbs, adjectives and adverbs). We impose no constraints on the part-of-speech category of the synsets where the word and its translation need to be found. If, for example, *world* and its French translation *monde* are found in both nominal and adjectival synsets, we extract all variants available in the synsets. This adds flexibility to the matching given that a word of a certain grammatical category might be translated by a word of a different category in another language.

The WSD method enriches each reference sentence with semantic variants valid in this precise context. For example, variants provided for the sentence: *Only healthcare workers allowed in*, include {exclusively, solely, alone, ...}, {health care practitioner, healthcare provider, health care professionals, ...}, {let, permit}. The disambiguation might fail, especially in cases where alignment information is not available or cannot be used because of the limited coverage of the BabelNet resource in languages other than English. When the annotations are correct, they help the Meteor metric reward translations in the hypothesis that are different from the ones in the reference but still semantically correct.

2.5 Results

Our results are reported using Kendall's τ for segment-level evaluation and Pearson's correlation coefficient for system-level evaluation, all computed with the official scripts and human judgments provided by the WMT14 shared metrics task organizers. The *xx* column in the results tables shows the average of all the language pairs involved. [6]

The results of Meteor-WSD at the segment-level are reported in Table 1. Meteor-WSD correlates slightly better with human judgments than standard Meteor when English is the target language, with an average improvement of .001. The results are also better than the results obtained by our previous version of Meteor-WSD (Apidianaki and Marie, 2015), especially for the cs-en language pair where correlation goes from .278 to .282. The differences between Meteor and

[5]The merge would lead to errors only in cases of parallel ambiguities where the word and its translation carry the same distant senses. Using translations in multiple languages could improve accuracy in these cases.

[6]This means that the score given for xx-en is the average of the scores of all language pairs with English as a target language. For xx-xx, the score is the average of all scores for all language pairs.

Metric	fr-en	de-en	cs-en	ru-en	xx-en	en-fr	en-de	en-cs	en-ru	en-xx	xx-xx
Meteor-1.5	.406	.334	.282	.329	.338	.280	.238	.318	.427	.316	.327
Meteor-WSD	.410	.332	.282	.332	.339	.280	.240	.321	.437	.320	.330

Table 1: Segment-level Kendall's τ correlations of Meteor-WSD and the official WMT14 human judgments.

Metric	fr-en	de-en	cs-en	ru-en	xx-en	en-fr	en-de	en-cs	en-ru	en-xx	xx-xx
Meteor-1.5	.975	.927	.980	.805	.922	.941	.263	.976	.923	.776	.849
Meteor-WSD	.975	.927	.979	.828	.927	.946	.258	.981	.929	.779	.852

Table 2: System-level Pearson's coefficient correlations of Meteor-WSD and the official WMT14 human judgments.

Meteor-WSD scores are much larger when English is the source language, probably due to the fact that we activate the synonymy module[7] and perform disambiguation in the other languages using the semantic information provided in BabelNet while Meteor uses synonyms only for English. This means that the synonyms left after disambiguation in languages other than English are useful and help to improve the correlation with human judgments. Table 2 presents our results at the system-level. As for the segment-level task, Meteor-WSD performs better than Meteor for almost all language pairs, with a significant improvement of .023 for the ru-en language pair.

3 A Metric Combination: RATATOUILLE

3.1 The Metrics

RATATOUILLE is a metric combination involving ten metrics mainly dedicated to segment-level evaluation: PER, WER, CDER (Leusch et al., 2006), TER (Snover et al., 2006), GTM 1.3 (Melamed et al., 2003), sentence-level BLEU, Meteor 1.5, Meteor-WSD, RIBES 1.03.1 (Echizen'ya et al., 2013) and BEER 1.0 (Stanojević and Sima'an, 2014). For the metrics PER, WER, CDER, TER and sentence-level BLEU we used the implementations available in MOSES (Koehn et al., 2007). For the metrics RIBES[8] and BEER[9] we used the implementations published by their authors, and the implementation available in the Asiya toolkit[10] (Giménez and Màrquez, 2010) for the GTM metric.

3.2 Tuning

Each metric of the combination gives a score for the evaluated segment. The score computed by RATATOUILLE is the result of the log-linear combination of each metric's score. The weight for each metric score is tuned using a similar approach to PRO (Hopkins and May, 2011), already used by Guzmán et al. (2014) in the context of metric combination evaluation. In this pairwise approach, candidate translation pairs are classified into two categories: correctly or incorrectly ordered, reducing the tuning to a binary classification problem. We studied two configurations, retaining all possible translation pairs or only pairs including translations separated by at least three ranks in the human judgments. We follow PRO which uses only pairs of translations of significant different quality and does not learn to tease apart translations of similar quality. Translation pairs used to tune the metric for a given language pair include translations in the same target language independently of the source language. If no human judgments are available for a given language pair, we use all the translation pairs independently of the target and source languages to tune the metric.[11] The classifier used is a MaxEnt from the scikit-learn python library (Pedregosa et al., 2011).

[7] As the synonymy module has no pre-defined weight for such translation directions, we tuned its weight on the WMT13 human judgments for each translation direction, searching empirically for the best weight between 0 and 1 with a 0.2 step size.

[8] http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html

[9] https://github.com/stanojevic/beer/

[10] http://nlp.lsi.upc.edu/asiya/

[11] For the fi-en language pair in the WMT15 metrics task, we used translation pairs from xx-en to tune the metric for fi-en and from en-xx to tune the metric for en-fi.

RATATOUILLE tuning set	fr-en	de-en	cs-en	ru-en	xx-en	en-fr	en-de	en-cs	en-ru	en-xx	xx-xx
all	.426	.336	.294	.337	.348	.292	.286	.352	.459	.347	.348
>=3	.425	.342	.297	.340	.351	.293	.292	.345	.456	.347	.349

Table 3: Segment-level Kendall's τ correlations of RATATOUILLE and the official WMT14 human judgments using all WMT13 human judgments (all) or only all the translation pairs containing translations separated by at least 3 ranks (>=3).

Metric	fr-en	de-en	cs-en	ru-en	xx-en	en-fr	en-de	en-cs	en-ru	en-xx	xx-xx
BEER	.417	.337	.284	.333	.343	.292	.268	.344	.440	.336	.340
RATATOUILLE w/o Meteor-WSD	.423	.343	.296	.338	.350	.293	.291	.344	.454	.346	.348
RATATOUILLE w/o Meteor-1.5	.425	.341	.297	.339	.351	.293	.292	.345	.458	.347	.349
RATATOUILLE	.425	.342	.297	.340	.351	.293	.292	.345	.456	.347	.349

Table 4: Segment-level Kendall's τ correlations of RATATOUILLE and the official WMT14 human judgments.

Metric	fr-en	de-en	cs-en	ru-en	xx-en	en-fr	en-de	en-cs	en-ru	en-xx	xx-xx
Meteor 1.5	.975	.927	.980	.805	.922	.941	.263	.976	.923	.776	.849
RATATOUILLE w/o Meteor-WSD	.974	.900	.994	.804	.918	.955	.403	.979	.946	.821	.869
RATATOUILLE w/o Meteor-1.5	.974	.899	.993	.804	.918	.958	.408	.979	.945	.823	.870
RATATOUILLE	.974	.901	.993	.804	.918	.959	.408	.979	.944	.823	.870

Table 5: System-level Pearson's coefficient correlations of RATATOUILLE and the official WMT14 human judgments.

3.3 Results

To tune RATATOUILLE, we used only the human judgments provided at WMT13.[12] As shown by Joty et al. (2014), using more data brings no improvements when tuning metric combinations. For system-level scores, the RATATOUILLE score for each sentence is first passed through a sigmoid function[13] and the final system score is the average of all sentence scores.

In the first experiments with RATATOUILLE, we tried to find a better subset of tuning examples among all the WMT13 translation pairs. We present in Table 3 our results when tuning on all translation pairs or on a subset including only translation pairs separated by at least three ranks in the human judgments. In spite of an important reduction in the number of translation pairs used to tune, we observed slight improvements in the average for xx-en, from .348 to .351, while the average for en-xx remains the same. We assume that this is probably due to the small number of translation pairs remaining for tuning after filtering; these are far less numerous for language pairs with English as source language than for language pairs with English as target language. Since on average the translation pair filtering gives better results, we report results for our experiments where we used the >=3 subsets to tune RATATOUILLE.

The results obtained for RATATOUILLE at the segment-level are presented in Table 4 along with the results of BEER, the best metric among the metrics that participated in the WMT14 metrics task for all language pairs. RATATOUILLE gives significantly better results than BEER – as expected, since BEER is used by RATATOUILLE – with an average improvement of .009. The largest improvements are observed for en-de (+.024) and en-ru (+.016). For en-fr and en-cs, RATATOUILLE results are only slightly better than BEER results (+.001), meaning probably that BEER is not assisted by the other metrics in RATATOUILLE to improve correlation with human judgments.

BEER did not participate in the WMT14 system-level evaluation. Meteor participated in this evaluation for all language pairs, so in Table 5 we present the RATATOUILLE results along with the results for Meteor. At this level, RATATOUILLE performs better than Meteor but

[12] http://www.statmt.org/wmt13/results.html

[13] We found out that not converting the scores with a sigmoid function leads to a slightly lower correlation. Indeed without this conversion scores are not bounded and can be very different between sentences especially for long sentences for which scores are very high, giving them more weight when computing the average for the system-level score.

not for all language pairs. We observe, for instance, a loss of .026 for de-en while we notice a strong improvement of .145 for en-de. This confirms the difficulty to have consistent results across language pairs at the system level as shown in the official results of the WMT14 metrics task where only one metric (PER) performed best on more than one translation directions, en-cs and en-ru, while different metrics perfomed best for each of the remaining en-xx translation directions.

For both segment and system levels, we also observed that withdrawing Meteor-1.5 from RATATOUILLE does not change the results on average, while withdrawing Meteor-WSD slightly decreases RATATOUILLE performance. This means that Meteor-WSD can successfully replace Meteor-1.5 in RATATOUILLE giving slightly better results.

4 Conclusion

We have shown the positive impact brought by introducing a word sense disambiguation step in an MT evaluation metric. Although lexical variation was addressed in previous metrics such as Meteor and Meteor-NEXT, synonyms and paraphrases were considered without taking the actual context into account. The improved correlation of the Meteor-WSD metric to human judgments of translation quality confirms the important role of the context in sense and synonym selection. The performance of the disambiguation method remains a crucial factor determining the performance of the MT evaluation metric. In future work, we intend to experiment with ways of improving disambiguation quality and increasing its coverage. Moreover, we intend to integrate context-based filtering of paraphrases to help the Meteor-WSD metric establish better matches between the compared translations. Last but not least, as BEER uses Meteor to align hypotheses and reference translations, we plan to replace Meteor by Meteor-WSD in BEER to improve this alignment and produce a better correlation with human judgments than the original BEER metric.

5 Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions. This work was partly supported by ANR project Transread (ANR-12-CORD-0015).

References

Marianna Apidianaki and Li Gong. 2015. LIMSI: Translations as Source of Indirect Supervision for Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, Denver, Colorado, USA.

Marianna Apidianaki and Benjamin Marie. 2015. METEOR-WSD: Improved Sense Matching in MT Evaluation. In *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 49–51, Denver, Colorado, USA, June.

Marianna Apidianaki, Emilia Verzeni, and Diana Mccarthy. 2014. Semantic Clustering of Pivot Paraphrases. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, USA.

Colin Bannard and Chris Callison-Burch. 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, June.

Helena Medeiros de Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2):59–77.

Michael Denkowski and Alon Lavie. 2010. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 250–253, Los Angeles, California, USA.

Hiroshi Echizen'ya, Kenji Araki, and Eduard Hovy. 2013. Automatic evaluation metric for machine translation that is independent of sentence length. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 230–236, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Jesús Giménez and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, pages 687–698, Baltimore, Maryland, USA, June.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1352–1362, Stroudsburg, PA, USA.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2014. Discotk: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT'14)*, pages 402–408, Baltimore, Maryland, USA, June.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL, demo session*, Prague, Czech Republic.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. Cder: Efficient mt evaluation using block movements. In *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*.

Dan I. Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*, NAACL-Short '03, pages 61–63, Stroudsburg, PA, USA.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research*, volume 12, pages 2825–2830.

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

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, Cambridge, USA.

Miloš Stanojević and Khalil Sima'an. 2014. BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, Maryland, USA, June.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June. Association for Computational Linguistics.