METEOR-WSD: Improved Sense Matching in MT Evaluation

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

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

Abstract

We present an initial experiment in integrating a disambiguation step in MT evaluation. We show that accounting for sense distinctions helps METEOR establish better sense correspondences and improves its correlation with human judgments of translation quality.

1 Introduction

Synonym and paraphrase support are useful means for capturing lexical variation in Machine Translation evaluation. In the METEOR metric (Banerjee and Lavie, 2005), some level of abstraction from the surface forms of words is achieved through the "stem" and "synonymy" modules which map words with the same stem or belonging to the same WordNet synset (Fellbaum, 1998). METEOR-NEXT (Denkowski and Lavie, 2010) extends semantic mapping to languages other than English and to longer text segments, using the paraphrase tables constructed by the *pivot* method (Bannard and Callison-Burch, 2005). Although both metrics yield improvements regarding correlation with human judgments of translation quality compared to the standard METEOR configuration for English, they integrate semantic information in a rather simplistic way: matching is performed without disambiguation, which means that all the variants available for a particular text fragment are treated as semantically equivalent. This is however not always the case, as synonyms found in different WordNet synsets correspond to different senses. Similarly, paraphrase sets obtained by the pivot method often group phrases describing different senses (Apidianaki et al., 2014). In these cases, a word sense disambiguation (WSD) step would help to identify the correct synset or subset of paraphrases for a word or phrase in context and avoid erroneous matchings between text segments carrying different senses. We present an initial experiment on the integration of a disambiguation step in the METEOR metric and show how it helps increase correlation with human judgments of translation quality.

2 Disambiguation in METEOR

We apply the metric to translations of news texts from the five languages involved in the WMT14 Metrics Shared Task (Machacek and Bojar, 2014) (French, Hindi, German, Czech, Russian) into English. We disambiguate the English references – different for each language pair – using the Babelfy tool (Moro et al., 2014), which performs graph-based WSD by exploiting the structure of the multilingual network BabelNet (Navigli and Ponzetto, 2012). The assigned annotations are multilingual synsets grouping word and phrase variants in different languages coming from various sources (WordNet, Wikipedia, etc.) and carrying the same sense. We use the WordNet literals found in the sense selected by Babelfy to filter the WordNet synonym sets used in METEOR and prevent METEOR from considering erroneous matchings as correct.[1] As a result, only the synonyms found in the proposed BabelNet synset are kept and considered as correct by METEOR, while synonyms corresponding to other

[1] In future work, we intend to apply the same filtering to paraphrases in different languages.

Proceedings of SSST-9, Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 49–51,
Denver, Colorado, June 4, 2015.

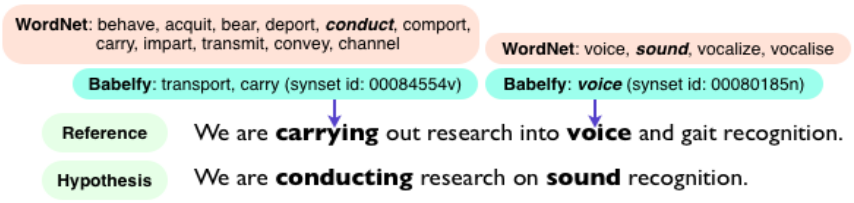

Figure 1: Good and erroneous matchings made by the synonymy module and WSD.

senses are discarded. METEOR is a tunable metric able to assign a weight to each of its modules in order to better correlate with human judgments. Since METEOR needs to perform a costly grid-search on 8 parameters, we did not re-optimize the weights due to time constraints. Considering this, the following experiments are made in a suboptimal configuration as we can expect a re-optimization to take the impact of the disambiguation into account more efficiently.

3 Results

In Table 1, we present the results obtained for four different configurations: METEOR with WordNet synonym support vs METEOR with WSD, with and without paraphrasing. The scores correspond to segment-level Kentall τ correlations of the metric with human judgments of translation quality. When the paraphrase module is activated, WSD slightly improves the correlation of the metric to human judgments in all languages except for Czech. Nevertheless, it is worth noting that this would improve METEOR's ranking in the results of the WMT14 shared task for French-English, which would then be ranked 4th, instead of 7th, among 18 participants.

When the WSD prediction is correct, it permits to avoid erroneous matchings between synonyms corresponding to different WordNet senses. In the example given in Figure 1, the synonymy module creates a wrong mapping between *sound* and *voice*. As *sound* is not contained in the BabelNet synset selected by the WSD component, this avoids establishing an erroneous match. Given, however, that WSD does not always succeed, the paraphrase module manages to find correspondences in cases of wrong disambiguation choices. This is the case illustrated by the first annotation in Figure 1 where the synset proposed by the WSD tool describes the "transport" sense. This wrong WSD prediction establishes no match but the paraphrase module that operates after WSD, manages to map *carrying* and *conducting*. When the paraphrase module is deactivated, the correlation of METEOR-WSD is lower than that of the basic METEOR configuration. Although the disambiguation discards erroneous matchings made by the synonymy module, there is no means to correct erroneous disambiguation choices without the paraphrases.

METEOR configuration		fr-en	de-en	hi-en	cs-en	ru-en
w/ par.	METEOR	.406	.334	.420	**.282**	.329
	METEOR-WSD	**.410**	**.335**	**.422**	.278	**.331**
w/o par.	METEOR	.400	.326	.401	.271	.313
	METEOR-WSD	.403	.321	.396	.263	.312

Table 1: Segment-level Kendall's τ correlations between METEOR and the official human judgments of the WMT14 metrics shared task.

4 Conclusion and Perspectives

Our results demonstrate the beneficial impact of disambiguation in MT evaluation. Accounting for sense distinctions helps METEOR establish better quality correspondences between hypotheses and human references. In future work, we intend to experiment with other WSD methods such as the alignment-based method recently proposed by Apidianaki and Gong (2015). Moreover, we plan to integrate a WSD step in evaluation for languages other than English. We expect to observe substantial improvements in languages where the synonymy module is unavailable and where the quality of pivot paraphrases is lower than in English. We also plan to conduct experiments using METEOR-WSD for tuning a Statistical Machine Translation system and expect to observe improvements in translation quality compared to the same system tuned with METEOR without WSD.

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, 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, USA.

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.

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

Matous Machacek and Ondrej Bojar. 2014. Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, 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.