METEOR-WSD: Improved Sense Matching in MT Evaluation

Marianna Apidianaki[1] *& Benjamin Marie*[1,2]
(1) LIMSI-CNRS, Orsay, France (2) Lingua et Machina, Le Chesnay, France

Lexical Variation in MT

METEOR (Banerjee and Lavie, 2005)

- stem and synonymy modules: mapping of words with the same stem or belonging to the same WordNet synset

METEOR-NEXT (Denkowski and Lavie, 2010)

- semantic mapping extended to languages other than English and to longer text segments using pivot paraphrases (Bannard and Callison-Burch, 2005)

Pros and Cons

+ Increased correlation with human judgments

- better matches compared to standard METEOR configuration

– Sense matching without WSD

- all available variants are treated as semantically equivalent
- synonyms found in different WordNet synsets correspond to different senses and pivot paraphrases often describe different senses

Why WSD?

- identify the correct synset or paraphrase subset for a word/phrase in context
- avoid erroneous matchings between text fragments carrying different senses

Data sets and tools

- English translations of news texts from the five languages of the **WMT14 Metrics Shared Task**: French, Hindi, German, Czech, Russian (Machacek and Bojar, 2014)
- English references disambiguated using **Babelfy**, a graph-based WSD tool that exploits the structure of the multilingual network Babelnet (Navigli and Ponzetto, 2012, Moro et al., 2014)

Disambiguation procedure

- **Babelfy annotations**: multilingual synsets grouping word and phrase variants in different languages coming from various sources (WordNet, Wikipedia, etc.) and carrying the same sense

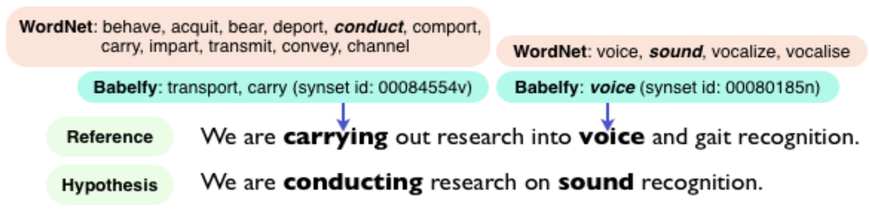
⇒ synonyms in the BabelNet synset selected by the WSD component are kept and considered as correct by METEOR

⇒ synonyms corresponding to other senses are discarded

- WSD prevents considering erroneous matchings as correct
- experiments carried out in a suboptimal configuration: METEOR re-optimization is expected to take the impact of WSD into account more efficiently

METEOR synonymy module vs WSD matchings

- METEOR synonymy module creates a wrong mapping between sound and voice
- WSD component prevents establishing an erroneous match: sound not in the selected synset
- the performance of the WSD method is very important
- when WSD fails, the paraphrase module manages to find correspondences



Experimental Results

- Segment-level Kendall's τ correlations between Meteor and the official human judgments of the WMT14 metrics shared task

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

Conclusion

- **WSD has a beneficial impact in MT evaluation**: accounting for sense distinctions helps METEOR establish better correspondences between hypotheses and references
- future work
  - experiment with other WSD methods (Apidianaki and Gong, SemEval-2015)
  - integrate WSD in evaluation for languages other than English
  - context-based filtering of pivot paraphrases
  - use METEOR-WSD for tuning an SMT system