**METEOR-WSD: Improved Sense Matching in MT Evaluation**

Marianna Apidianaki¹ & Benjamin Marie¹,²

(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)

---

**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)

---

**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

---

**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

---

**Experimental Results**

<table>
<thead>
<tr>
<th>METEOR configuration</th>
<th>fr-en</th>
<th>de-en</th>
<th>hi-en</th>
<th>cs-en</th>
<th>ru-en</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/ par.</td>
<td>METEOR</td>
<td>.406</td>
<td>.334</td>
<td>.420</td>
<td>.282</td>
</tr>
<tr>
<td></td>
<td>METEOR-WSD</td>
<td>.410</td>
<td>.335</td>
<td>.422</td>
<td>.278</td>
</tr>
<tr>
<td>w/o par.</td>
<td>METEOR</td>
<td>.400</td>
<td>.326</td>
<td>.401</td>
<td>.271</td>
</tr>
<tr>
<td></td>
<td>METEOR-WSD</td>
<td>.403</td>
<td>.321</td>
<td>.396</td>
<td>.263</td>
</tr>
</tbody>
</table>

**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