Integrating Multi Word Expressions in Statistical Machine Translation

Zied elloumi1, Olivier Kraif2, Laurent Besacier3

1) LIG-CETALD, France
2) Universite Grenoble Alpes - LIDILEM, France
3) Firstname.Lastname@u-grenoble3.fr

Abstract

Multiword Expressions (MWE) are complex phenomena which represent a real challenge for computational linguistics, and Statistical Machine Translation systems (SMT) in particular. In this paper, we try to improve translation of Multiword Expressions (MWE) are complex phenomena which represent a real challenge for computational linguistics, and Statistical Machine Translation systems (SMT) in particular. In this paper, we try to improve translation of Multiword Expressions (MWE). Preliminary results, regarding the BLEU score, are encouraging. }

MT and MWEs

MWEs are a challenge for MT:

- Example 1: Idiom
  - Source EN: my job was a piece of cake
  - Google TR: mon travail était un morceau de gâteau

- Example 2: Phrasal verbs
  - Source EN: [...] turned the stereo on
  - Google TR: [...] touma la stéréo sur

Pre-processing of PV for En-Fr SMT training

- Identification of PVs of our test corpus using the parts of speech and some dependencies provided by XIP [Aït-Mokhtar et al., 2002]
- Gluing verb and prep for training and test (Tst-MWE)
- Examples:
  - keep your voice down, Hermione begged him.
  - He picked up a remote and turned the stereo on.

Processing of idioms

- Using a pre-defined « dictionary » of idioms
- Constrained SMT decoding [koehn,2014]
- Example:
  - And you thought my job was a piece of cake!
  - BLEU goes from 15.21 % to 30.71 % on the 134 sentences containing idioms

Extraction of a specific corpus for evaluating MT of MWEs

- Create a List of MWEs based on electronic dictionaries available on the Net, Forums
- Using LIDILEM’s tool - Lexico-gramme [O. Kraif et S. Divervy (2012)] - Europarl, News and Enolex (literary texts) corpora are indexed

Results

<table>
<thead>
<tr>
<th>Corpus of 500 sent.</th>
<th>Detail per corpus or per category on Tst-MWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus of 500 sent.</td>
<td>Tst-NORMAL</td>
</tr>
<tr>
<td>Moses-LIG</td>
<td>24.87%</td>
</tr>
<tr>
<td>Google-TR</td>
<td>23.81%</td>
</tr>
</tbody>
</table>

BLEU scores (Moses-LIG vs Google)

- Slight but not significant improvement
- Need an evaluation metric that could “zoom” on MWEs
- BLEU [Papineni and al., 2002] is calculated on the whole corpus → do not see the effect on smaller events

Conclusion

- Corpora with bigger density of MWEs than in usual text
- Need for an automatic metric that focuses on specific events such as MWEs (otherwise we do not see much differences)
- First propositions to handle MWE
  - Pre-processing of PV for En-Fr SMT
  - Use constrained decoding for Idioms
- Example:
  - Source: surely they must call the operation off now?
  - Hyp (baseline): ils doivent annuler l’opération.
  - Hyp (pre-proc): ils doivent annuler le fonctionnement maintenant?

References