Instance selection improves cross-lingual model training for fine-grained sentiment analysis

July 30th, 2015

Roman Klinger and Philipp Cimiano
roman.klinger@ims.uni-stuttgart.de

IMS, University of Stuttgart and CITEC, Bielefeld University
CoNLL 2015
It looks very neat, like a storage container and using it is very simple and easy.

- Training data is available in one language but not in another:
  - How can we automatically translate and project?
  - What is the performance?
  - Can we improve by instance filtering with translation quality estimation?
Es gibt mit Sicherheit **bessere** Maschinen, aber die bietet das **beste** Preis-Leistungs-Verhältnis.

There are certainly **better** machines, but offers the **best** price-performance ratio.
Methods

- **Model**: Supervised probabilistic model for joint aspect and evaluating phrase detection
- **Translation**: Google Translate API
- **Alignment**: FastAlign
- **Projection**: Shortest match including all tokens aligned with an annotation
- **Filtering**: Based on machine translation quality estimation:
  - Language model for source language
  - Language model for target language
  - Likelihood of alignment
Results Teaser for Aspects

- **In-target-language Training**: 41% $F_1$ measure
- **Projection**: 23% $F_1$ measure
- **Filtering**: 47% $F_1$ measure
Instance selection improves cross-lingual model training for fine-grained sentiment analysis

Roman Klinger and Philipp Cimiano

Summary

Motivation: Scarcity of annotated corpora for many languages is a bottleneck for training fine-grained sentiment analysis models that can tag aspects and subjective phrases.

Challenge: Statistical machine translation and projecting annotated data from a source language to a target language supports building a resource for new languages, but quality may be limited when training on that resource: Performance drops from 41% F1 to 23% F1 for aspects.

Idea: Removing low-quality translations by filtering instances maintains quality: Performance of up to 41% F1 for aspect phrases. Translation of subjective phrases is less challenging.

Motivation

• Sentiment Analysis/Opinion Mining are important for a lot of domains
• Annotated corpora are mainly available for English
• Our goal: Automatically building annotated resources by machine translation and annotation projection which enable supervised training of models with performance competitive to in-target-language training

Research Questions

• What is the performance on the task when...
  – ... training data for the source language is projected into a target language
  – ... when training data for the target language is available?
• Can the performance be increased by selecting high-quality translations?

Methods

Model

• Probabilistic model to phrase detection based on surface features and dependency parsing
• MCMC inference for coupled prediction of evaluating phrases and aspect phrases
• No prior knowledge in addition to training corpus
• Implementation available based on FACTORIE

Machine Translation and Projection

• Open Source Tool (e.g. Moses SMT)
  – Choice of parallel training corpus difficult: EuroParl only mentions few relevant concepts
  – Instead: Google Translate and alignment as postprocessing with FastAlign
• Projection transfers annotation to the shortest phrase in the target language which contains all tokens in the source language annotation

Quality Estimation and Filtering

• Idea: Do not use all instances but only the ones which are “good” – similar to real language. We use three SMT quality measures:
  1. Source language probability based on language model
  2. Target language probability based on language model
  3. Likelihood of alignment based on FastAlign

Experiments

Data

• USAGE Corpus for German and English
• Corpus of Amazon Reviews for different products in two languages
• Sentence-wise manual annotation of quality for all translations de→en
• Cross-domain evaluation: Train on six product categories and test on one
• Test on manually annotated data in target language

Different Thresholds (de→en)

<table>
<thead>
<tr>
<th>Threshold</th>
<th>German to English, Aspects</th>
<th>German to English, Subjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td><img src="chart1.png" alt="Chart" /></td>
<td><img src="chart2.png" alt="Chart" /></td>
</tr>
<tr>
<td>Alignment</td>
<td><img src="chart3.png" alt="Chart" /></td>
<td><img src="chart4.png" alt="Chart" /></td>
</tr>
</tbody>
</table>

Contact: Roman Klinger, roman.klinger@ims.uni-stuttgart.de