Alignment-Guided Chunking

Yanjun Ma, Nicolas Stroppa, Andy Way

{yma,nstroppa,away}@computing.dcu.ie

National Center for Language Technology
Dublin City University

TMI 2007
Outline

Motivation

Alignment-Guided Chunking
  Definition
  Remarks

Experimental Results
  Data
  Chunking Results
  Application

Conclusion & Future work
motivation

monolingual V.S. bilingual context

- word segmentation V.S. word alignment
  - tokenize the source and target language in bilingual context
    (Ma et al. 2007)
- chunk up sentences in bilingual context?
motivation

different sentence chunking for EBMT

- Example-based Machine Translation
  - English-to-French translation
  - English-to-German translation
  - we should chunk English differently!

SMT decoding

- log-linear phrase-based SMT (Och & Ney, 2002)

\[
\log P(e^l_1 | f^l_1) = \sum_{m=1}^{M} \lambda_m h_m(e^l_1, f^l_1) + \lambda_{LM} \log P(e^l_1)
\]
motivation

SMT decoding

- log-linear phrase-based SMT

\[ \log P(e_1|f_1^J) = \sum_{m=1}^{M} \lambda_m h_m(e_1, f_1^J, s^K_1) + \lambda_{LM} \log P(e_1), \quad (2) \]

where \( s^K_1 = s_1...s_k \) denotes a segmentation of the source and target sentences respectively into the sequence of phrases \((\tilde{e}_1, ..., \tilde{e}_k)\) and \((\tilde{f}_1, ..., \tilde{f}_k)\)

- in decoding, \( s^K_1 \) is not usually modeled, meaning the context of the source language is missing (see Stroppa et al., 2007)
a chunking model with following features

- predict the chunking pattern of a given sentence in a bilingual context
- adaptable to different end-tasks, i.e. different language pairs in MT
- integration into state-of-the-art EBMT & SMT systems
motivation

monolingual chunks

- marker-based chunks (Gough & Way, 2004; Stroppa & Way, 2006)

bilingual chunks

- IBM fertility models (Brown et al., 1993)
- joint probability model (Marcu & Wong, 2002; Burch et al., 2006)
- semi-supervised bilingual chunking (Liu et al., 2004)
- ITG (Wu, 1997)
monolingual chunking in bilingual context

<table>
<thead>
<tr>
<th></th>
<th>data</th>
<th>goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNLL</td>
<td>monolingual; manually crafted</td>
<td>shallow parsing (linguistically motivated)</td>
</tr>
<tr>
<td>marker</td>
<td>monolingual; manually crafted</td>
<td>chunk alignment for MT</td>
</tr>
<tr>
<td>semi-supervised</td>
<td>bilingual; no word alignment</td>
<td>chunk alignment for MT</td>
</tr>
<tr>
<td>ITG</td>
<td>bilingual; word alignment</td>
<td>bilingual parsing</td>
</tr>
<tr>
<td>AGC</td>
<td>bilingual; word alignment</td>
<td>monolingual chunking for MT</td>
</tr>
</tbody>
</table>
Outline

Motivation

Alignment-Guided Chunking
  Definition
  Alignment-Guided Chunking
  Remarks

Experimental Results
  Data
  Chunking Results
  Application

Conclusion & Future work
alignment-guided chunking: definition

- bilingual corpus

Cette ville est chargée de symboles puissants pour les trois religions monothéistes.
The city bears the weight of powerful symbols for all three monotheistic religions.

- word alignment

0-0 1-1 2-2 3-4 4-5 5-7 6-6 7-8 8-9 9-10 10-12 11-11 12-13

- alignment-guided chunks

Cette ville est chargée de symboles puissants pour les trois religions monothéistes.
The city bears the weight of powerful symbols for all three monotheistic religions.
Alignment-Guided Chunking

**main idea**

*learn chunking model from bilingual corpus*

- chunks are learned from bilingual corpus
- all the information learned can be re-used in machine translation

**steps**

- use a word aligner to align words
- derive alignment-guided chunks for source language using word alignment
- estimate a probabilistic model for *monolingual* chunking
- chunk new sentences
Alignment-Guided Chunking

Data representation

Data representation for CoNLL-style chunks

- **IOB1, IOB2, IOE1, IOE2, IO, ]**, [ (Tjong Kim Sang & Veenstra, 1999)

Our data representation scheme

- **IB** - all chunk-initial words receive a B tag
- **IE** - all chunk-final words receive a E tag
- **IBE1** - all chunk-initial words receive a B tag, all chunk-final words receive a E tag; if there is only one word in the chunk, it receives a B tag
- **IBE2** - all chunk-initial words receive a B tag, all chunk-final words receive a E tag; if there is only one word in the chunk, it receives a E tag
Alignment-Guided Chunking

**parameter estimation**

**feature selection**
- words and their POS tags

**machine learning techniques**
- maximum entropy (Berger et al., 1996; Koeling, 2000)
- memory-based learning (Daelemans & Van den Bosch, 2005)
Remarks

a new look at chunking

Figure: example of alignment-guided chunking

- make hard decision for each word to get a chunked sentence
- transform chunking from a binary classification task into a ranking task
- provide more information for end-tasks
Outline

Motivation

Alignment-Guided Chunking
  Definition
  Remarks

Experimental Results
  Data
  Chunking Results
  Application

Conclusion & Future work
data and preprocessing

**Europarl corpus**

- French-English and German-English
- focus on English chunking
- training set: around 300k aligned sentences sharing the same English sentences
- test set: 21,972 sentence pairs (1 reference)
- tools: Giza++ (Och & Ney, 2003) for word alignment, MXPOST (Ratnaparkhi, 1996) for POS tagging, maxent (Zhang, 2004) and TiMBL (Daelemans et al. 2007) for discriminative chunking
Data

statistics on training data

<table>
<thead>
<tr>
<th></th>
<th>English-French</th>
<th>English-German</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of Chunks</td>
<td>3,316,887</td>
<td>2,915,325</td>
</tr>
<tr>
<td>shared chunks [%]</td>
<td>42.08</td>
<td>47.87</td>
</tr>
</tbody>
</table>

Table: number of chunks in English sentences for different bilingual corpus

- average English chunk length - 1.84 words for French-English corpus and 2.10 words for German-English corpus
- chunking model should vary from task to task
**Motivation**

**Alignment-Guided Chunking**

**Experimental Results**

**Conclusion & Future work**

---

**Chunking Results**

**results - alignment-guided chunking**

*(German-to-English)*

<table>
<thead>
<tr>
<th></th>
<th>accuracy</th>
<th>precision</th>
<th>recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaxEnt</td>
<td>68.41</td>
<td>47.57</td>
<td>35.12</td>
<td>40.41</td>
</tr>
<tr>
<td>MBL</td>
<td>65.75</td>
<td>38.00</td>
<td>41.61</td>
<td>39.72</td>
</tr>
</tbody>
</table>

**Table:** alignment-guided chunking results

- both the precision and recall are low, even the accuracy
- maximum entropy performs better on precision, but worse on recall
- contexts are too complicated and could be inconsistent
- voting techniques using different models
speeding SMT by filtering translation table (German-to-English)

<table>
<thead>
<tr>
<th>t-table size</th>
<th>BLEU [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBSMT</td>
<td>4,765,052</td>
</tr>
<tr>
<td>AGC filter</td>
<td>1,019,697</td>
</tr>
<tr>
<td>random filter</td>
<td>1,019,697</td>
</tr>
</tbody>
</table>

**Table:** influence of translation table filtering

- might help when time and space are limited
- related work (Johnson et al., 2007)
## Outline

**Motivation**

**Alignment-Guided Chunking**
- Definition
- Remarks

**Experimental Results**
- Data
- Chunking Results
- Application

**Conclusion & Future work**
conclusion

- propose a new approach - alignment-guided chunking, for monolingual chunking in bilingual context
- a probabilistic model that can be used to model source sentence segmentation in SMT decoding (see section 1)
- use different machine learning techniques for alignment-guided chunking
- prove to be effective for t-table filtering in SMT
- potential use in log-linear phrase-based SMT
discussion

- disadvantage - mismatch between training and testing
  - training
    - make use of bilingual information
    - word alignment and chunking are two separate processes
  - testing - monolingual information
- advantage - mismatch between training and testing
  - perform sentence chunking in bilingual context
future work

- evaluate the model in a log-linear phrase-based SMT system
- evaluate the model in EBMT system
- parameter estimation - test different features and feature combinations
- use multi-reference to evaluate the chunking results
Thank you for listening
NULL words

- check the following words - W NULL or W W
- never partition - NULL W or NULL NULL
configuration of machine learning toolkits

- maximum entropy
  - parameter estimation - default. Limited-Memory Variable Metric (L-BFGS)
- memory-based learning
  - parameter estimation - default. IB1, weighted overlap
Filtering t-table in SMT

- given a phrase pair, check the context of the specific phrase
- the leftmost word *preceding* the phrase should be a chunk-final word
- the rightmost word *inside* this phrase should be a chunk-end word