MaTrEx: DCU Machine Translation System for IWSLT 2006

Nicolas Stroppa and Andy Way

Dublin City University, School of Computing, NCLT
Outline

DCU@IWSLT 2006

System’s description

Results
DCU@IWSLT 2006

First Participation

- Open Data Track
- Two directions:
  - Italian → English
  - Arabic → English
- 1-best ASR hypotheses + correct recognition results
- Use the provided training data only
Outline

DCU@IWSLT 2006

System’s description

Results
MaTrEx: A Hybrid EBMT/SMT System

Overview of the system

- A word alignment component (GIZA++)
- A chunking component
- A chunk alignment component
- Two phrase alignment components:
  - “SMT”-style phrase aligner (standard phrase extraction from GIZA++ alignments)
  - “EBMT”-style phrase aligner (phrases are extracted from (i) the chunker and (ii) the chunk aligner)
- A minimum-error rate training component (Phramer)
- A decoder (Pharaoh)
- A case and punctuation restoration component
Chunking

Two types of chunking

- Marker-based chunking
  - surface chunking based on marker words
- Treebank-based chunking
  - learner trained on annotated data extracted from treebanks
Marker-Based Chunking

- Approach to EBMT based on the Marker Hypothesis
  
  "The Marker Hypothesis states that all natural languages have a closed set of specific words or morphemes which appear in a limited set of grammatical contexts and which signal that context." (Green, 1979).

- Universal psycholinguistic constraint: languages are marked for syntactic structure at surface level by closed set of lexemes or morphemes.
Marker-Based Chunking

In my case, it is usually on business, seldom for pleasure.
Marker-Based Chunking

In my case, it is usually on business, seldom for pleasure.

- NPs usually start with determiners, or possessive pronouns
- PPs usually start with prepositions
Marker-Based Chunking

In my case, it is usually on business, seldom for pleasure.

Nel mio caso, solitamente per affari, raramente per piacere.

- NPs usually start with determiners, or possessive pronouns
- PPs usually start with prepositions
Marker-Based Chunking

In my case, it is usually on business, seldom for pleasure.
Nel mio caso, solitamente per affari, raramente per piacere.

- NPs usually start with determiners, or possessive pronouns
- PPs usually start with prepositions
- We can use a set of closed-class marker words to segment aligned source and target sentences (determiners, quantifiers, prepositions, conjunctions, possessive pronouns, personal pronouns, punctuation marks)
Treebank-based Chunking

Chunking as a sequence tagging task

Example (using the Inside-Outside-Begin representation)

It takes time to train a train driver .
PRP V NN TO V DT NN NN .
B-NP B-VP B-NP B-VP I-VP B-NP I-NP I-NP O
[ ]_{NP} [ ]_{VP} [ ]_{NP} [ ]_{VP} [ ]_{NP}

- Tagged data can be extracted from Treebanks (cf. CoNLL 2000 shared task)
- Sequence tagging is performed using a classifier (sliding window)
- Efficient classifiers such as Support Vector Machines (SVM) can be applied (implemented in e.g. \textsc{Yamcha})
Chunking

- The lists of English and Italian marker words were extracted from Celex and MorphIt respectively, and edited manually.
- We trained Yamcha on the English and Arabic Penn Tree Banks.
Chunk Alignment

The chunks obtained from the chunkers have to be aligned

**English:** [it felt okay] [after the game] [but then] [it started turning black-and-blue] [is it serious ?]
**Italian:** [era a posto] [dopo la partita] [ma poi] [ha cominciato] [a diventare livida] [è grave ?]

**English:** [in my case] [it is usually] [on business] [seldom] [for pleasure]
**Italian:** [nel mio caso] [solitamente] [per affari] [raramente] [per piacere]
Chunk Alignment Strategies

We assume that for a pair of aligned chunked sentences \((e^k_1, f^l_1)\), we have access to \(P(e_i|f_j)\) and \(P(f_j|e_i)\).

Several alignment strategies

- Edit-distance-like alignment
- Edit-distance-(with jumps)-like alignment
- *IBM model-1-like alignment*

We perform the alignments in both directions and keep the intersection.
Chunk Alignment

How to compute $P(e_i|f_j)$?

We can use the following information:

- Marker Tags
- Cognate Information
- Word Translation Probabilities (IBM model 1-like)

We can combine the different sources of knowledge within a log-linear model

Remark
All the parameters are computed “on the fly”.
Combining the “SMT” and “EBMT” chunks

Hybridity

- EBMT and SMT aligned chunks are merged (counts are added)

Adding EBMT chunks to the SMT chunks:

- adds good alignments which are not present otherwise (less constrained strategy than the phrase extraction heuristic)
- “boosts” already present SMT chunks, i.e. contributes to the direct re-estimation of phrases (rare phrases are over-estimated)
Outline

DCU@IWSLT 2006

System’s description

Results
Data and Preprocessing

- Training was performed using the provided data (no external data)
- We used the OpenNLP tokenizer (a Maximum-Entropy approach) for tokenizing the English and Italian data (a set of regular expressions was added for Italian), and \texttt{Asvm} for Arabic
- Chunking was done with the marker-based chunker (English, Italian), and \texttt{Asvm} (Arabic)
- Chunk Alignment was performed using the Edit-Distance-like aligner (Italian$\rightarrow$English), and the Edit-Distance-with-jumps-like aligner (Arabic$\rightarrow$English)
- 3-gram Language Model, with Kneser-Ney smoothing
- Minimum-Error Rate Training on dev4 dataset
- Punctuation and case information was restored using \texttt{Srilm} (hidden-ngram and disambig)
- Removed the words in the output that were directly copied from the input
Results - Arabic

**ASR (1-best)**

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>Meteor</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official</td>
<td>0.145</td>
<td>4.531</td>
<td>0.402</td>
<td>0.7027</td>
<td>0.5949</td>
</tr>
<tr>
<td>Additional</td>
<td>0.1391</td>
<td>4.794</td>
<td>0.4</td>
<td>0.7165</td>
<td>0.5870</td>
</tr>
</tbody>
</table>

**Correct Recognition Result**

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>Meteor</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official</td>
<td>0.1624</td>
<td>4.89</td>
<td>0.4336</td>
<td>0.686</td>
<td>0.5678</td>
</tr>
<tr>
<td>Additional</td>
<td>0.1589</td>
<td>5.29</td>
<td>0.432</td>
<td>0.6935</td>
<td>0.5537</td>
</tr>
</tbody>
</table>
## Results - Italian

### ASR (1-best) - Baseline

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>Meteor</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Official</strong></td>
<td>0.2399</td>
<td>6.39</td>
<td>0.5378</td>
<td>0.60</td>
<td>0.49</td>
</tr>
<tr>
<td><strong>Additional</strong></td>
<td>0.2568</td>
<td>6.98</td>
<td>0.54</td>
<td>0.59</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Results - Italian

ASR (1-best) - Baseline (B) vs. Matrex (M)

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>Meteor</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official-B</td>
<td>0.2399</td>
<td>6.39</td>
<td>0.5378</td>
<td>0.60</td>
<td>0.49</td>
</tr>
<tr>
<td>Official-M</td>
<td>0.2598</td>
<td>6.59</td>
<td>0.5497</td>
<td>0.5835</td>
<td>0.4869</td>
</tr>
<tr>
<td>Additional-B</td>
<td>0.2568</td>
<td>6.98</td>
<td>0.54</td>
<td>0.59</td>
<td>0.46</td>
</tr>
<tr>
<td>Additional-M</td>
<td>0.2783</td>
<td>7.228</td>
<td>0.5495</td>
<td>0.5662</td>
<td>0.4498</td>
</tr>
</tbody>
</table>

About 2 BLEU points improvement
Results - Italian

<table>
<thead>
<tr>
<th>Correct Recognition Result - Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BLEU</strong></td>
</tr>
<tr>
<td>Official</td>
</tr>
<tr>
<td>Additional</td>
</tr>
</tbody>
</table>
## Results - Italian

### Correct Recognition Result - Baseline (B) vs. Matrex (M)

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>Meteor</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Official-B</strong></td>
<td>0.2882</td>
<td>7.223</td>
<td>0.6219</td>
<td>0.5528</td>
<td>0.43548</td>
</tr>
<tr>
<td><strong>Official-M</strong></td>
<td>0.3126</td>
<td>7.546</td>
<td>0.6246</td>
<td>0.5315</td>
<td>0.4286</td>
</tr>
<tr>
<td><strong>Additional-B</strong></td>
<td>0.3219</td>
<td>8.046</td>
<td>0.6220</td>
<td>0.5188</td>
<td>0.3788</td>
</tr>
<tr>
<td><strong>Additional-M</strong></td>
<td>0.3467</td>
<td>8.358</td>
<td>0.6245</td>
<td>0.4964</td>
<td>0.3744</td>
</tr>
</tbody>
</table>

More than 2 BLEU points improvement
Summary

- We introduced the MaTrEx Data-Driven MT system being developed at Dublin City University
- We presented a method to extract aligned phrases using chunkers and chunk aligners:
  - Marker-based chunking, SVM-based chunking
  - Edit-Distance-like chunk aligners
- We participated in the OpenData Track, for the Italian-to-English and Arabic-to-English directions
Ongoing and Future Work

- Perform a more systematic comparison of the different chunking and alignment strategies
- Insert a segmentation probability directly in the decoding process, in order to give a preference to the phrases that are chunks according to the chunker
- Insert chunk label information in a factored model
Thank you for your attention

http://www.computing.dcu.ie/research/nclt