MaTrEx: Machine Translation Using Examples

Nicolas Stroppa and Andy Way

NCLT, School of Computing, Dublin City University

DCU NCLT @ NIST MT 2006
Outline

1 Background

2 System’s description

3 Results/Discussion
1 Background

2 System’s description

3 Results/Discussion
National Centre for Language Technology (NCLT) in DCU. A team of 12 researchers:
- 2 M.Sc Students, 7 Ph.D. Students, 2 Postdocs
- Supervised by Dr. Andy Way

First Participation to NIST MT. In 2006:
- OpenLab (TC STAR), Spanish → English
- NIST MT, Arabic → English
- IWSLT, Arabic → English, Italian → English

Large-scale Example-Based Machine Translation system
- Easily adaptable to new language pairs
- Modular design - follow established Design Patterns
- Hybrid system: EBMT/SMT
Historically, we have been working on EBMT
EBMT and SMT are showing more and more similarities (use of aligned “phrases”)
We are working more and more on the combination of EBMT and SMT resources
Some problems and mistakes

- Strong underestimation of the workload: only one person, part-time, for 5 weeks
- Problems with memory requirement (> 4 Gigs of RAM needed by Giza++)
- Main cluster unavailable for 3 days because of maintenance during the last week
- Buckwalter had been automatically lowercased (!!)
- LMs were not trained on English GigaWord (only UN Data)
- MERT was skipped, EBMT chunking and alignment were skipped!

⇒ the results do not reflect the capabilities of our system!
Background

SYSTEM’S DESCRIPTION

Results/Discussion
**A Phrase-Based EBMT/SMT System**

- **Data-driven system:** Makes use of aligned phrases extracted from sententially-aligned corpora
- **Two types of extraction:**
  - “SMT” phrases extracted from words alignments (GIZA++ + heuristic)
  - “EBMT” phrases extracted thanks to (i) a chunking and (ii) an alignment of chunks proposed by the EBMT system
### A Phrase-Based EBMT/SMT System

- Data-driven system: Makes use of aligned phrases extracted from sententially-aligned corpora.
- Two types of extraction:
  - "SMT" phrases extracted from words alignments (Giza++ + heuristic)
  - "EBMT" phrases extracted thanks to (i) a chunking and (ii) an alignment of chunks proposed by the EBMT system.
**Matrex: A Hybrid EBMT/SMT System**

**A Phrase-Based EBMT/SMT System**

- Data-driven system: Makes use of aligned phrases extracted from sententially-aligned corpora
- Two types of extraction:
  - “SMT” phrases extracted from words alignments (GIZA++ + heuristic)
  - “EBMT” phrases extracted thanks to (i) a chunking and (ii) an alignment of chunks proposed by the EBMT system
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.

The Dearborn Mich., energy company stopped paying a dividend in the third quarter of 1984 because of troubles at its Midland nuclear plant.
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.

The Dearborn Mich., energy company stopped paying a dividend in the third quarter of 1984 because of troubles at its Midland nuclear plant
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.

3 NPs start with determiners, one with a possessive pronoun
- Determiners & possessive pronoun - small closed-class sets
- Predicts head nominal element will occur in the right-context.

The Dearborn Mich. energy company stopped paying a dividend in the third quarter of 1984 because of troubles at its Midland nuclear plant
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.

3 NPs start with determiners, one with a possessive pronoun
- Determiners & possessive pronoun - small closed-class sets
- Predicts head nominal element will occur in the right-context.

Four prepositional phrases, with prepositional heads
- Again a small set of closed-class words
- Indicates that soon thereafter an NP object will occur

The Dearborn Mich., energy company stopped paying a dividend in the third quarter of 1984 because of troubles at its Midland nuclear plant
Use a set of closed-class marker words to segment aligned source and target sentences during a pre-processing stage.

- **<PUNC>** used as end of chunk marker

<table>
<thead>
<tr>
<th>Marker Type</th>
<th>Marker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determiner</td>
<td>&lt;DET&gt;</td>
</tr>
<tr>
<td>Quantifiers</td>
<td>&lt;Q&gt;</td>
</tr>
<tr>
<td>Prepositions</td>
<td>&lt;P&gt;</td>
</tr>
<tr>
<td>Conjunctions</td>
<td>&lt;C&gt;</td>
</tr>
<tr>
<td>WH-Adverbs</td>
<td>&lt;WH&gt;</td>
</tr>
<tr>
<td>Possessive Pronouns</td>
<td>&lt;POSS-PRON&gt;</td>
</tr>
<tr>
<td>Personal Pronouns</td>
<td>&lt;PERS-PRON&gt;</td>
</tr>
<tr>
<td>Punctuation Marks</td>
<td>&lt;PUNC&gt;</td>
</tr>
</tbody>
</table>

English Marker words extracted from CELEX and edited manually.
Use a set of closed-class marker words to segment aligned source and target sentences during a pre-processing stage.

- <PUNC> used as end of chunk marker

<table>
<thead>
<tr>
<th>Category</th>
<th>Marker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determiner</td>
<td>&lt;DET&gt;</td>
</tr>
<tr>
<td>Quantifiers</td>
<td>&lt;Q&gt;</td>
</tr>
<tr>
<td>Prepositions</td>
<td>&lt;P&gt;</td>
</tr>
<tr>
<td>Conjunctions</td>
<td>&lt;C&gt;</td>
</tr>
<tr>
<td>WH-Adverbs</td>
<td>&lt;WH&gt;</td>
</tr>
<tr>
<td>Possessive Pronouns</td>
<td>&lt;POSS-PRON&gt;</td>
</tr>
<tr>
<td>Personal Pronouns</td>
<td>&lt;PERS-PRON&gt;</td>
</tr>
<tr>
<td>Punctuation Marks</td>
<td>&lt;PUNC&gt;</td>
</tr>
</tbody>
</table>

- English Marker words extracted from CELEX and edited manually.
Marker-Based EBMT: Chunking (3)

Pros

- Psycho-Linguistic motivation
- Simple (linear)
- Easily adaptable (only a list of marker-words is needed)
- Does not need expensive training on treebanks, etc.

Cons

- Blind (no context taken into account)
- Deterministic
- Not so easily adaptable to languages such as Arabic/Chinese (POS would be needed) ⇒ we used ASVM for Arabic chunking

Remarks

- Can be combined with different chunkers, e.g. machine-learning based chunkers (cf. CoNLL’2000 shared task)
- In the English PTB, the most frequent first words of chunks are mostly marker-words…
“Edit-Distance Like” Chunk Alignment. Does not depend on the chunking strategy.
- Dynamic programming algorithm
- Conditional probabilities used:
  - Based on Marker Tags
  - Based on Cognate Information: \( \text{Lowest Common Subsequence Ratio, Dice Coefficient, Minimum Edit-Distance} \)
  - Based on Word Translation Probabilities
  - Combination (\( \Rightarrow \) can be viewed as a log-linear model)

\[
\lambda_1 d_1(a, b) + \cdots + \lambda_n d_n(a, b) \Rightarrow -\lambda_1 \log P_1(a|b) \cdots - \lambda_n \log P_n(a|b)
\]

- “Edit-Distance” with Jumps
  - Useful for languages where the word order is different (didn't improve results for Spanish/English MT)
“Edit-Distance Like” Chunk Alignment. Does not depend on the chunking strategy.

- Dynamic programming algorithm

Conditional probabilities used:

- Based on Marker Tags
- Based on Cognate Information: Lowest Common Subsequence Ratio, Dice Coefficient, Minimum Edit-Distance
- Based on Word Translation Probabilities
- Combination (⇒ can be viewed as a log-linear model)

\[
\lambda_1 d_1(a, b) + \cdots + \lambda_n d_n(a, b) \Rightarrow -\lambda_1 \log P_1(a|b) \cdots - \lambda_n \log P_n(a|b)
\]

“Edit-Distance” with Jumps

- Useful for languages where the word order is different (didn’t improve results for Spanish/English MT)
“Edit-Distance Like” Chunk Alignment. Does not depend on the chunking strategy.

- Dynamic programming algorithm

Conditional probabilities used:

- Based on Marker Tags
- Based on Cognate Information: Lowest Common Subsequence Ratio, Dice Coefficient, Minimum Edit-Distance
- Based on Word Translation Probabilities
- Combination (⇒ can be viewed as a log-linear model)

\[ \lambda_1 d_1(a, b) + \cdots + \lambda_n d_n(a, b) \Rightarrow -\lambda_1 \log P_1(a|b) \cdots - \lambda_n \log P_n(a|b) \]

“Edit-Distance” with Jumps

- Useful for languages where the word order is different (didn’t improve results for Spanish/English MT)
**Chunk Alignment**

- "Edit-Distance Like" Chunk Alignment. Does not depend on the chunking strategy.
  - Dynamic programming algorithm
- Conditional probabilities used:
  - Based on Marker Tags
  - Based on Cognate Information: *Lowest Common Subsequence Ratio*, *Dice Coefficient*, *Minimum Edit-Distance*
  - Based on Word Translation Probabilities
  - Combination (\( \Rightarrow \) can be viewed as a log-linear model)

\[
\lambda_1 d_1(a, b) + \cdots + \lambda_n d_n(a, b) \Rightarrow -\lambda_1 \log P_1(a|b) \cdots - \lambda_n \log P_n(a|b)
\]

- "Edit-Distance” with Jumps
  - Useful for languages where the word order is different (didn’t improve results for Spanish/English MT)
“Edit-Distance Like” Chunk Alignment. Does not depend on the chunking strategy.

- Dynamic programming algorithm
- Conditional probabilities used:
  - Based on Marker Tags
  - Based on Cognate Information: *Lowest Common Subsequence Ratio, Dice Coefficient, Minimum Edit-Distance*
  - Based on Word Translation Probabilities
  - Combination (⇒ can be viewed as a log-linear model)

\[ \lambda_1 d_1(a, b) + \cdots + \lambda_n d_n(a, b) \Rightarrow -\lambda_1 \log P_1(a|b) \cdots - \lambda_n \log P_n(a|b) \]

“Edit-Distance” with Jumps

- Useful for languages where the word order is different (didn’t improve results for Spanish/English MT)
**Chunk Alignment**

- "Edit-Distance Like" Chunk Alignment. Does not depend on the chunking strategy.
  - Dynamic programming algorithm
- Conditional probabilities used:
  - Based on Marker Tags
  - Based on Cognate Information: *Lowest Common Subsequence Ratio*, *Dice Coefficient*, *Minimum Edit-Distance*
  - Based on Word Translation Probabilities
  - Combination (⇒ can be viewed as a log-linear model)

\[
\lambda_1 d_1(a, b) + \cdots + \lambda_n d_n(a, b) \Rightarrow -\lambda_1 \log P_1(a|b) \cdots - \lambda_n \log P_n(a|b)
\]

- "Edit-Distance" with Jumps
  - Useful for languages where the word order is different (didn’t improve results for Spanish/English MT)
Mixing chunks

Hybridity

- “EBMT” and “SMT” aligned chunks are merged
- Adding EBMT chunks to the SMT chunks database:
  - adds good alignments which are not present otherwise
  - “boosts” already present SMT chunks (re-estimation)
Other tools

- **Pre-processing**
  - English: OpenNLP. Sentence segmentation and tokenization
  - Arabic: ASVM. Tokenization

- **Part-of-Speech Tagging**
  - English: TreeTagger
  - Arabic: ASVM

- **Chunking**
  - English: Marker-Based chunking/SVM chunking (Yamcha)
  - Arabic: ASVM

- **Note:** nothing done with dates, names, etc.
**System Architecture**

- **Input sentence**
  - Aligned Sentences
  - Decoder
  - Output translation

- **Word aligner**
  - Aligned Words

- **Chunk aligner**
  - Aligned Chunks
  - Source marker words
  - Target marker words

- **Source chunks**
  - Target chunks

- `'distance' metric`
Aligned Sentences are submitted to word alignment and chunk alignment modules to produce translation resources

- Modular in design
Aligned Sentences are submitted to word alignment and chunk alignment modules to produce translation resources

- Modular in design
**System Architecture**

- Aligned Sentences are submitted to word alignment and chunk alignment modules to produce translation resources.
- Modular in design.
- Easily adaptable and extendible.
Outline

1. Background
2. System’s description
3. Results/Discussion
What do these results mean? Virtually nothing (they are those of a broken SMT system)

Do not reflect the system’s capability

Admitted failure to scale. Wanted to play the game anyway.
Plan to continue the development the MaTrEx system
  - Currently at early stage of development
  - Implement an HMM-based chunk alignment strategy
  - Investigate better the implication of hybridity
  - Implement an Example-Based decoder (i.e. strong prior on chunking) + Use of generalised templates
  - Big improvement expected for NIST MT 2007…
Thank you for your attention.

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