Integrating corpus-based and rule-based approaches in an open-source machine translation system

Felipe Sánchez-Martínez Juan Antonio Pérez-Ortiz
Mikel L. Forcada

Transducens Group – Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain

{fsanchez,japerez,mlf}@dlsi.ua.es

New Approaches to Machine Translation. A METIS-II Workshop
Leuven, Belgium
January 11, 2007

Outline

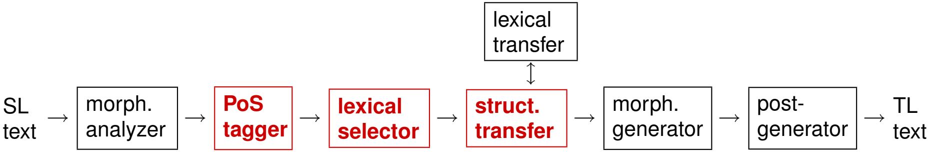
1. Apertium
   - Open-source shallow-transfer MT engine
   - On the use of target language information
2. Target-language-driven part-of-speech tagger training
   - Introduction
   - Method overview
3. Target-language-driven lexical selector training
   - Introduction
   - Method overview
4. Shallow-transfer rules extraction from parallel corpora
   - Introduction
   - Method overview
5. Discussion

Outline

1. Apertium
   - Open-source shallow-transfer MT engine
   - On the use of target language information
2. Target-language-driven part-of-speech tagger training
   - Introduction
   - Method overview
3. Target-language-driven lexical selector training
   - Introduction
   - Method overview
4. Shallow-transfer rules extraction from parallel corpora
   - Introduction
   - Method overview
5. Discussion

Open-source shallow-transfer MT engine

http://apertium.org



- Engine completely independent from the linguistic data
- Linguistic data coded using XML-based formats
- Compilers convert linguistic data into an efficient form

On the use of TL information to train SL models

lexical transfer

↕

SL text → morph. analyzer → **PoS tagger** → **lexical selector** → struct. transfer → morph. generator → post-generator → TL text

- Use of information from the target side of the MT system, and two monolingual corpora in both languages
- Some modules of the MT system are used to train the remaining modules

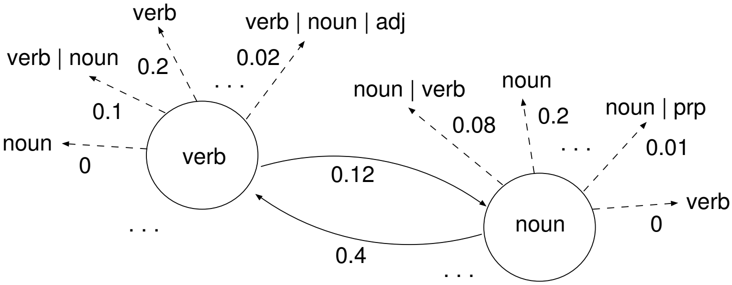
Outline

1. Apertium
   - Open-source shallow-transfer MT engine
   - On the use of target language information
2. Target-language-driven part-of-speech tagger training
   - Introduction
   - Method overview
3. Target-language-driven lexical selector training
   - Introduction
   - Method overview
4. Shallow-transfer rules extraction from parallel corpora
   - Introduction
   - Method overview
5. Discussion

Part-of-speech tagging

- Problem: Selecting the correct part-of-speech for those words with more than one (ambiguous words)

- Part-of-speech tagger based on first-order hidden Markov models (HMM)



Target-language-driven part-of-speech tagger training

- Unsupervised training
- The method uses the rest of the modules of the MT system in which the resulting tagger will be embedded
- The SL corpus is segmented and for each segment all possible disambiguations are translated into TL
- A TL model is used to choose the best disambiguations
- HMM parameters are computed according to the likelihood of the corresponding translations into TL

$\Rightarrow$ The resulting tagger is tuned to the translation quality $\Leftarrow$

Example (English→Spanish)

- SL sentence (English):
  - He–`prn` books–`noun|verb` the–`art` room–`noun|verb`
- Possible translations (Spanish) according to each disambiguation and their normalized likelihoods according to a TL model:

• Él–`prn` reserva–`verb` la–`art` habitación–`noun`	0.75
• Él–`prn` reserva–`verb` la–`art` aloja–`verb`	0.15
• Él–`prn` libros–`noun` la–`art` habitación–`noun`	0.06
• Él–`prn` libros–`noun` la–`art` aloja–`verb`	+ 0.04
	1.00

- The HMM parameters involved in these 4 disambiguations are updated according to their likelihoods in the TL

Latest results

- Reported by Sánchez-Martínez, Pérez-Ortiz and Forcada (2006)
- TL model: A classical trigram language model
- Error rates calculated over ambiguous words only:
  - Supervised training $\simeq$ 11%
  - TL-driven (unsupervised) $\simeq$ 25%
  - Baum-Welch expectation maximization (unsupervised) $\simeq$ 31%

Outline

1. Apertium
   - Open-source shallow-transfer MT engine
   - On the use of target language information
2. Target-language-driven part-of-speech tagger training
   - Introduction
   - Method overview
3. Target-language-driven lexical selector training
   - Introduction
   - Method overview
4. Shallow-transfer rules extraction from parallel corpora
   - Introduction
   - Method overview
5. Discussion

Lexical selection

- Problem: Selecting the correct translation into TL for those words with more than one (ambiguous words)
- Lexical selector based on co-occurrence models of SL lemmas
- No TL information is used to perform the lexical selection during execution

Target-language-driven lexical selector training

- Use of two monolingual corpora and the bilingual dictionary of the MT system
- Use of a sliding context window and a list of *stopwords*
- Training procedure
  1. Each SL ambiguous word is translated into TL, and for each translation a co-occurrence model is built from the TL corpus
  2. Train a co-occurrence model of SL *translation senses*:
     - For each ambiguous SL word the set of possible translations is calculated
     - Co-occurrence information of TL lemmas is transferred to the SL by translating the surrounding context

Example (Spanish→English) /1

- SL context window (Spanish) after removing stopwords:

$$C(\text{gato}) = \{\text{arbol}, \text{perro}, \text{ladrar}\}$$

- Set of translation senses for SL lemma “gato”:

$$T(\text{gato}) = \{\text{gato}^{\text{jack}}, \text{gato}^{\text{cat}}\}$$

- Translated context (English):

$$C_t(\text{gato}) = \{\text{tree}, \text{dog}, \text{bark}\}$$

Example (Spanish→English) /2

- Scores found in the TL co-occurrence models are transferred to the SL models:

$$
\begin{array}{lcrcl}
S(\text{cat}, \text{tree}) & = & 100 & = & S(\text{gato}^{\text{cat}}, \text{arbol}) \\
S(\text{cat}, \text{dog}) & = & 300 & = & S(\text{gato}^{\text{cat}}, \text{perro}) \\
S(\text{cat}, \text{bark}) & = & 3 & = & S(\text{gato}^{\text{cat}}, \text{ladrar}) \\
S(\text{jack}, \text{tree}) & = & 12 & = & S(\text{gato}^{\text{jack}}, \text{arbol}) \\
S(\text{jack}, \text{dog}) & = & 23 & = & S(\text{gato}^{\text{jack}}, \text{perro}) \\
S(\text{jack}, \text{bark}) & = & 0 & = & S(\text{gato}^{\text{jack}}, \text{ladrar})
\end{array}
$$

Disambiguation method

- Disambiguation is performed in an analogous way
  - A sliding context window is considered
  - Given an ambiguous word, co-occurrence models for each translation senses are consulted
  - Scores for words in the context are added up
  - The translation sense with the higher final score becomes the winner
- Experimental results not yet available. We are currently working on it.

Outline

1 Apertium
- Open-source shallow-transfer MT engine
- On the use of target language information

2 Target-language-driven part-of-speech tagger training
- Introduction
- Method overview

3 Target-language-driven lexical selector training
- Introduction
- Method overview

4 Shallow-transfer rules extraction from parallel corpora
- Introduction
- Method overview

5 Discussion

Shallow-transfer rules extraction from parallel corpora

Transfer rules are used to:

- produce grammatically correct translations in the TL
- perform some lexical changes, such as preposition changes
- introduce auxiliary verbs when needed
- ...

Goal:

- To automatically learn those transformations that produce correct translations in the TL

How:

- Adapting the alignment templates (ATs) already used in statistical MT to the Apertium architecture

Method overview

Alignment templates are learned in a 3-stage procedure:

1. Compute word alignments
2. Extract aligned phrase pairs (translation units)
3. Generalize over the extracted phrases using word classes

Linguistic information used to define word classes:

- closed lexical categories: categories that cannot easily grow when adding new words to the dictionaries (pronouns, prepositions, etc.)

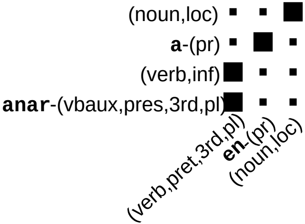
Word class: part-of-speech (including all inflection information)

- but closed words have their own single class

Alignment template example

Bilingual phrase:

Alignment template:



Spanish analysis:

vivieron en Alicante[1] $\longrightarrow$ *vivir*-`(verb,pret,3rd,pl)` **en**-`(pr)` *Alicante*-`(noun,loc)`

Catalan analysis:

van viure a Alacant $\longrightarrow$ **anar**-`(vbaux,pres,3rd,pl)` *viure*-`(verb,inf)` **a**-`(pr)` *Alacant*-`(noun,loc)`

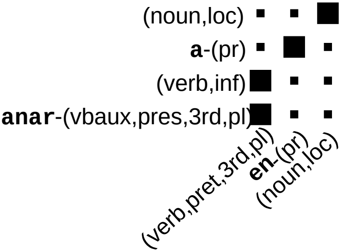
[1] Translated into English as *They lived in Alicante*

Alignment templates application

Spanish (input): *permanecieron en Alemania*[2] ⟶
permanecer-`(verb,pret,3rd,pl)` **en**-`(pr)`
Alemania-`(noun,loc)`

Catalan (output): **anar**-`(vbaux,pres,3rd,pl)`
romandre-`(verb,inf)` **a**-`(pr)`
Alemanya-`(noun,loc)` ⟶
van romandre a Alemanya

Word-for-word translation:
romangueren en Alemanya



[2] Translated into English as *They remained in Germany*

Latest results

- Reported by Sánchez-Martínez and Ney (2006)
- Use of a corpus with around 300 000 words for AT extraction
- Translation quality compared to that of word-for-word translation and that obtained with handcrafted rules
- Relative improvement if the quality achieved with handcrafted rules is assumed to be 100 % :
  - Spanish→Catalan: 70 %
  - Catalan→Spanish: 60 %

Outline

1. Apertium
   - Open-source shallow-transfer MT engine
   - On the use of target language information
2. Target-language-driven part-of-speech tagger training
   - Introduction
   - Method overview
3. Target-language-driven lexical selector training
   - Introduction
   - Method overview
4. Shallow-transfer rules extraction from parallel corpora
   - Introduction
   - Method overview
5. Discussion

Discussion

- Three corpus-based approaches have been presented to "give linguistic content" to some modules of the open-source (rule-based) MT system Apertium
- Some modules of the MT system itself are used to train the remaining modules
- Transfer rules are extracted from a small parallel corpus
  - Inferred transfer rules are human-readable. Automatically-inferred rules and handcrafted rules can coexist
- All methods are (or will shortly be) freely available under open-source licenses

Further reading

- Corbí-Bellot A. M. et al.
  An open-source shallow-transfer machine translation engine for the Romance languages of Spain
  Proceedings of the Tenth Conference of the EAMT, p. 79–80, 2005.

- Sánchez-Martínez F., J.A. Pérez-Ortiz and M. L. Forcada
  Speeding up target-language driven part-of-speech tagger training for machine translation
  LNAI 4293 (Advances in Artificial Intelligence, MICAI 2006), p. 844–854, 2006.

- Sánchez-Martínez F. and H. Ney
  Using alignment templates to infer shallow-transfer machine translation rules
  LNAI 4139 (Advances in Natural Language Processing, FinTAL 2006), p. 756–767, 2006.

Integrating corpus-based and rule-based approaches in an open-source machine translation system

Felipe Sánchez-Martínez Juan Antonio Pérez-Ortiz
Mikel L. Forcada

Transducens Group – Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain

{fsanchez,japerez,mlf}@dlsi.ua.es

New Approaches to Machine Translation. A METIS-II Workshop
Leuven, Belgium
January 11, 2007