Automatic induction of shallow-transfer rules for open-source machine translation

Felipe Sánchez-Martínez, Mikel L. Forcada

Transducens Group, Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
{fsanchez, mlf}@dlsi.ua.es

Introduction

Goal

- To automatically infer shallow-transfer rules, to be used in machine translation (MT), from "small" parallel corpora
- Transfer rules are used to:
  - produce grammatically correct translations in the target language (TL)
  - perform some lexical changes, such as preposition changes
  - introduce auxiliary verbs when needed
  - ...

How?

- Adapting the alignment templates already used in statistical MT

Resources

- A sentence-aligned parallel corpus
- A morphological analyzer and a PoS tagger for both languages (the ones used by the MT system in which the inferred rules will be used)

Alignment templates (AT)

- Introduced in the statistical MT framework as a feature function [1]
- Alignment templates (AT) 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
- AT $z = (S_n, T_m, A)$
  - S_n: sequence of n SL word classes
  - T_m: sequence of m TL word classes
  - A: alignment information

AT for shallow-transfer MT

- Linguistic information used to define word classes:
  - lexicalized categories: categories that are known to be involved in lexical changes such as prepositions
    - the method can learn not only syntactic changes
- Word class: part of speech with all the inflection information
  - but lexicalized words have their own single class

Extending ATs with restrictions

- ATs are extended to consider a set R of restrictions over the inflection information of non-lexicalized categories
  - AT $z = (S_n, T_m, A, R)$
- Restrictions are learned from the bilingual dictionary
  - Bilingual entry that does not change inflection information

```
<e><p>
  <l>castigo<s n="noun"/></l>
  <r>càstig<s n="noun"/></r>
</p></e>
```

R: w=**noun.***

  - Bilingual entry that does change inflection information

```
<e><p>
  <l>calle<s n="noun"/><s n="f"/></l>
  <r>carrer<s n="noun"/><s n="m"/></r>
</p></e>
```

R: w=**noun.m***

- The bilingual dictionary is also used to discard phrase pairs that cannot be reproduced by the MT system

Example of extracted ATs

Bilingual phrase:

estret
carrer
el
la
calle
estrecha

Alignment template:

(adj.m.sg)
(noun.m.sg)
el-(art.m.sg)
el(art.f.sg)
(noun.f.sg)
(adj.f.sg)

R= { w_2 =noun.m.*, w_3=adj.* }

Bilingual phrase:

Alacant
a
viure
van
vivieron en Alicante

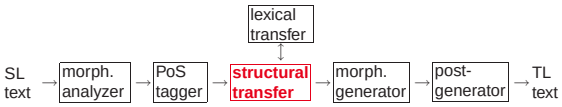
Alignment template:

(noun.loc)
a-(pr)
(verb.inf)
anar-(vbaux.pres.3rd.pl)
(verb.pret.3rd.pl)
el(pr)
(noun.loc)

R= { w_1 =verb.*, w_3=noun.* }

The Apertium open-source MT platform

http://www.apertium.org



Rules generation

- A shallow transfer rule consists of a set of ATs:

$$U = \{(S_m, T_n, A, R) \in Z : S_m = S^U\},$$

  where Z is the whole set of extracted ATs, and S^U is a sequence of SL word classes all ATs $z \in U$ have in common
- Each generated rule consists of code which always applies the most frequent AT $z = (S_m, T_n, A, R) \in U$ that satisfies the TL restrictions R
- A "default" AT, which translates word for word, is added with the lowest frequency

Code generated for each AT

- Code is generated for each unit in T_n, which depends on the type of word class:
  non-lexicalized word: the aligned SL (non-lexicalized) lemma is translated and inflection information provided by the TL word class is attached
  lexicalized word: it is introduced *as is*; it represents a complete lexical form
- Example:
  - Input: *vivir*-(verb.pret.3rd.pl) **en**-(pr) *Francia*-(noun.loc)
  - Output: **anar**-(vaux.pres.3rd.pl) *viure*-(verb.inf) **a**-(pr) *França*-(noun.loc)

AT applicability test

- Restrictions are tested by looking at the bilingual dictionary
- Example:

  - R = { w_2 =noun.m.*, w_3=adj.* }
  - Applicable:
    - Input string (Spanish): *la señal roja* ⟶ **el**-(art.f.sg) *señal*-(noun.f.sg) *rojo*-(adj.f.sg)
    - Translation of non-lexicalized words:
      - *señal*-(noun.f.sg)→*senyal*-(noun.m.sg)
      - *rojo*-(adj.f.sg)→*vermell*-(adj.f.sg)

  - Not applicable:
    - Input string (Spanish): *la silla blanca* ⟶ **el**-(art.f.sg) *silla*-(noun.f.sg) *blanco*-(adj.f.sg)
    - Translation of non-lexicalized words:
      - *silla*-(noun.f.sg)→*cadira*-(noun.f.sg)
      - *blanco*-(adj.f.sg)→*blanc*-(adj.f.sg)

Experiments (Spanish–Catalan)

- Lexicalized categories = { prep, pronoun, det, cnj, rel, vbmodal, vbaux }

Training corpus

Lang.	# sentences	# words
es	100 834	1 952 317
ca	100 834	2 032 925

Evaluation corpus

Trans. dir.	Corpus	# words
es-ca	post-edit	10 066
	parallel	13 147
ca-es	post-edit	10 024
	parallel	13 686

Results (WER)

Trans. dir.	Evaluation corpus	No rules	AT-based	Hand-coded
es-ca	post-edit	12.6 %	8.5 %	6.7 %
	parallel	26.6 %	20.4 %	20.8 %
ca-es	post-edit	11.6 %	8.1 %	6.5 %
	parallel	19.3 %	14.9 %	14.5 %

Discussion

- Significant improvement in translation quality as compared to word-for-word
- Translation quality very close to that obtained using hand-coded transfer rules
- Preliminary results on the Spanish–Portuguese language pair show results in agreement to those provided here
- Future work:
  - Applying shorter ATs inside the same rule when none of the longer ATs can be applied because of TL restrictions not being met
- An open.source implementation of the method can be freely downloaded from http://sf.net/projects/apertium/, package apertium-transfer-tools

References

[1] Och, F.J., H. Ney (2004). "The alignment template approach to statistical machine translation". In *Computational Linguistics*, 30(4):417–449.

[2] Armentano-Oller, C. et al. (2006). "Open-source Portuguese-Spanish machine translation". In *Lecture Notes in Computer Science 3960 (Computational Processing of the Portuguese Language)*, p. 50–59, Rio de Janeiro, Brazil.