Marker-Based Filtering of Bilingual Phrase Pairs for SMT

Felipe Sánchez-Martínez[†]

Andy Way[‡]

[†]Dept. Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant, Spain
fsanchez@dlsi.ua.es

[‡]NCLT, School of Computing
Dublin City University
Dublin 9, Ireland
away@computing.dcu.ie

14th May 2009; 13th Annual Meeting of the EAMT

Outline

Motivation & goal

Motivation:

- High number of bilingual phrase pairs extracted from a word-aligned sentence
  - All possible pairs within a certain *n*-gram length are considered
  - Number of pairs grows exponentially with the length of the sentences
- Large translation tables unmanageable for:
  - “on-demand” online machine translation
  - machine translation in mobile devices

Goal:

- Develop a simple method to filter bilingual phrase pairs
- Test the “Marker Hypothesis” in this tasks

Marker hypothesis

Marker Hypothesis : the syntactic structure of a language is *marked* at the surface level by a closed set of *marker* words (Green, 1979)

Successful in MT : Segmentation of aligned sentences into linguistically motivated bilingual chunks for EBMT

Haga click | en el botón rojo | para ver | la selección
Click | on the red button | to see | your selection

- Particularly useful to achieve good translation performance with small translation tables (Groves and Way, 2005ab)

$$\text{BLEU}_{\text{EBMT}} < \text{BLEU}_{\text{SMT}} < \text{BLEU}_{\text{EBMT+SMT}}$$

Marker-based filtering approach /1

Words are classified into two different categories:

Closed words : provide the structure for well-formed sentences; no special "intrinsic" meaning

- prepositions, pronouns, articles, ...
- no new words are usually added to this set as the language evolves

Open words : bear the meaning in a sentence

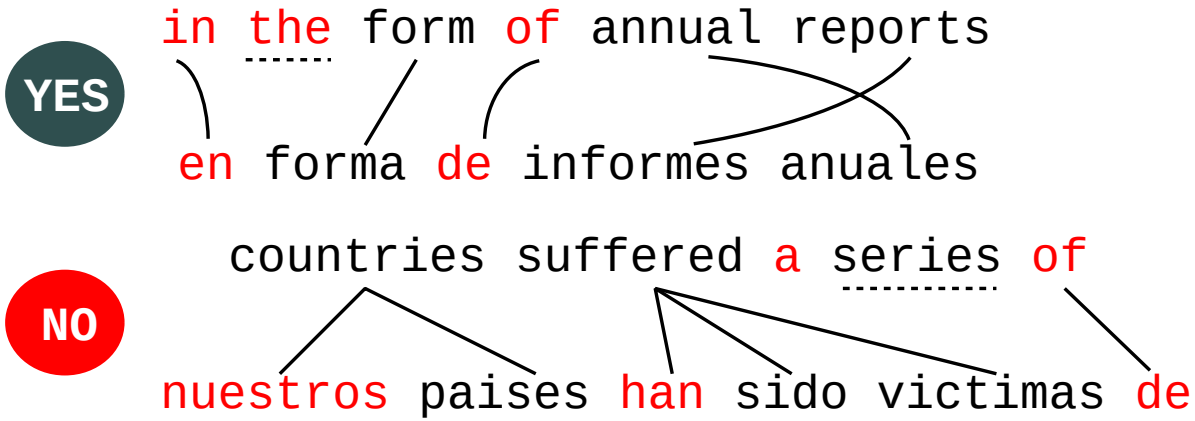
- nouns, verbs, adjectives, ...

Marker-based filtering approach /2

Filtering approach, underlying assumption:

- Accurate bilingual phrase pairs should have an alignment between the open words (bear meaning)

- Closed words may remain unaligned
  - the syntactic structure changes from one language to another

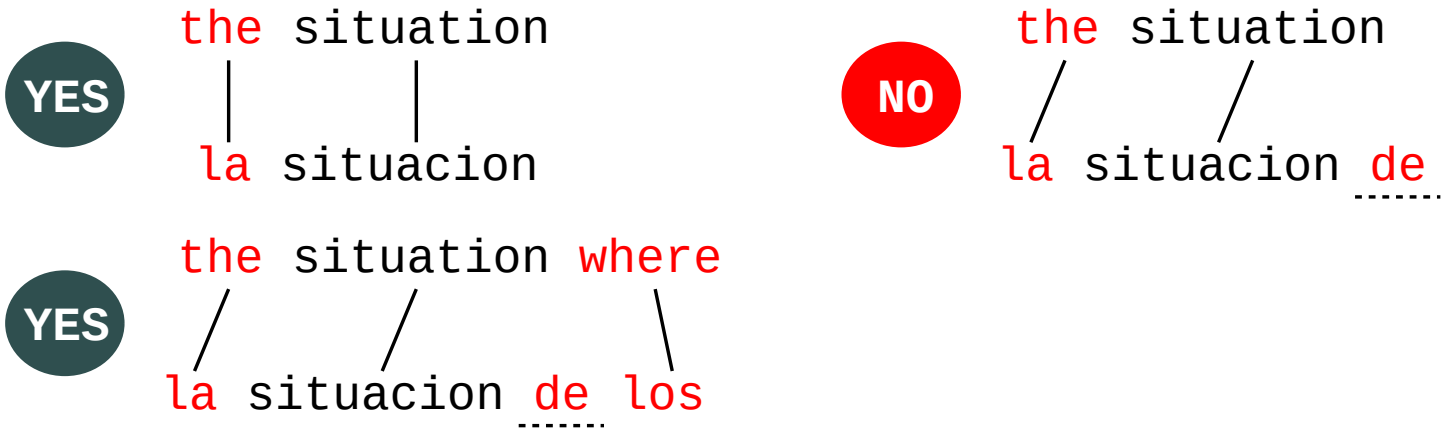


Marker-based filtering approach /3

Two different criteria to filter the bilingual phrase pairs:

"open words alig" : bilingual phrases containing one or more open words with no alignment are discarded

"open words alig+borders" : bilingual phrases are also discarded if the first or last word in either language has no alignment



Experimental setup /1

- Data distributed for the WMT 09 Workshop for MT
- Language pairs: `es-en`, `en-es`, `fr-en`, `en-fr`
- Using GIZA++, MOSES and the SRILM toolkit
- Bilingual phrases are filtered before scoring and running MERT

Experimental setup /2

Different lists of words used as marker words:

closed words : determiners, prepositions, pronouns, coordinate and subordinate conjunctions, relative and possessive pronouns, punctuation marks

Spanish:	<DET>: el, la, los, ...
(193 words)	<PREP>: de, para, ...
	<PRON>: yo, tú, él, ...
English:	<DET>: the, a, some, ...
(185 words)	<PREP>: on, it, up, ...
	<PRON>: you, he, she, ...
French:	<DET>: le, la, les, ...
(174 words)	<PREP>: sur, dans, par, ...
	<PRON>: vous, il, me, ...

Experimental setup /3

closed words+vaux : All inflected forms of auxiliary and modal verbs are also used

Spanish:	(1,572 words)	deber, haber, poder, querer, ser
English:	(201 words)	be, have
French:	(490 words)	avoir, devoir, être, falloir, pouvoir, vouloir

- Large number of words in Spanish due to verbs with attached enclitic pronouns

Experimental setup /4

stop words : the top *n* most frequent words found in the training corpora

Spanish:	de, la, que, en, el, y, a, los, ...
English:	the, to, of, and, a, in that, for, ...
French:	de, la, à, le, et, les, des, du, ...

- Number of stop words tested: 200, 100 and 50

Results /1

Spanish↔English

Pair	List of marker words	open words alig		+borders	
		filtered pairs	BLEU	filtered pairs	BLEU
es–en	baseline		0.2355		0.2355
	closed words	24.73%	0.2232	34.80%	0.2170
	closed words+vaux	23.72%	0.2188	34.69%	0.2157
	50 stop-words	31.63%	0.2090	41.15%	0.2037
en–es	baseline		0.2208		0.2208
	closed words	24.72%	0.2090	34.71%	0.2032
	closed words+vaux	23.69%	0.2112	34.59%	0.2039
	50 stop words	31.64%	0.2014	41.98%	0.1943

Results /2

French↔English

Pair	List of marker words	open words alig		+borders	
		filtered pairs	BLEU	filtered pairs	BLEU
fr-en	baseline		0.2331		0.2331
	closed words	33.04%	0.2128	41.26%	0.2072
	closed words+vaux	30.74%	0.2130	40.16%	0.2076
	50 stop words	35.14%	0.2082	44.31%	0.2029
en-fr	baseline		0.2105		0.2105
	closed words	33.08%	0.1965	41.20%	0.1928
	closed words+vaux	30.75%	0.1990	40.07%	0.1957
	50 stop words	35.18%	0.1903	44.24%	0.1885

Results /3

- Baseline system performs better than our approach in all cases
  - Difference is statistically significant according to the 95% confidence intervals (bootstrap resampling)
- Significant reduction of the phrase table at a reduced cost in translation performance
  - `es-en`, `en-es`: $\simeq$ 25% less phrases; BLEU $\simeq$ 0.012 worse
  - `fr-en`, `en-fr`: $\simeq$ 33% less phrases; BLEU $\simeq$ 0.017 worse
- Related work on this topic also show a small degradation of BLEU
  - But Jonshon et al. (2007) report a 90% reduction without worsening performance

Discussion /1

- A novel approach to filter bilingual phrases in SMT
- Tested on four European language pairs and with different lists of marker (closed) words
- May be useful in those cases in which a reduced system footprint is required
  - SMT integration in mobile devices

Discussion /1

Future work plan:

- Do not consider prepositions as closed words when they are part of a phrasal verb
- Test the approach for the translation from English to non-European languages such as Chinese, Japanese, or Hindi
  - “Maker Hypothesis” applies to any language

Marker-Based Filtering of Bilingual Phrase Pairs for SMT

Felipe Sánchez-Martínez[†]

Andy Way[‡]

[†]Dept. Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant, Spain
fsanchez@dlsi.ua.es

[‡]NCLT, School of Computing
Dublin City University
Dublin 9, Ireland
away@computing.dcu.ie

14th May 2009; 13th Annual Meeting of the EAMT