Relevance of Different Segmentation Options on Spanish-Basque SMT

Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola

Ixa Research Group
Euskal Herriko Unibertsitatea/Universidad del País Vasco
jipdisaa@ehu.es, gorka.labaka@ehu.es, jipsagak@ehu.es

BASQUE LANG.

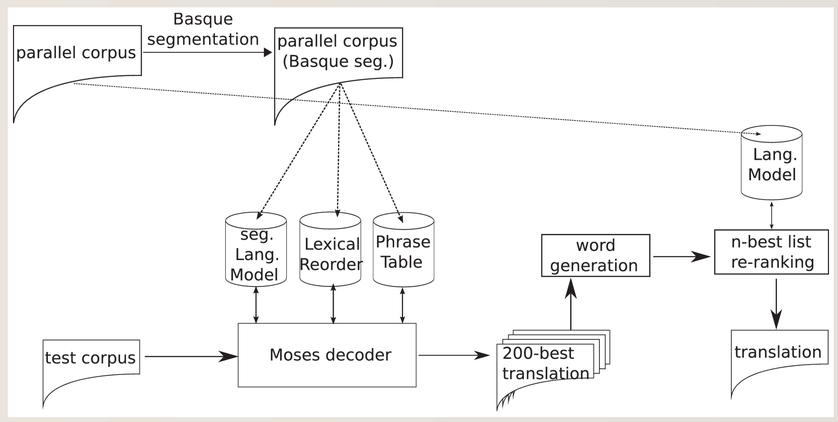
- Basque is a morphologically-rich language.
  - Translating into Basque requires a huge generation of morphological information.
  - Some Spanish words, like prepositions or articles, correspond to Basque suffixes. In case of ellipsis more than one suffix can be added to the same lemma.

etxe	etxeko	etxekoa	etxekoarengana
/house/	*/of the house/*	*/the one of the house/*	*/towards the one of the house/*

- Basque is free-order. The order of the constituents in the sentence is free.
  - Both, "Gizona etxera doa" and "etxera doa gizona" (*/the man goes home/*), are correct in Basque.
- Basque is a low-density language:
  - Basque is spoken by almost 1,000,000 people, but only 500,000 people it is their first language.
  - There are very few corpora available and they are smaller than corpora available for other languages.

SEGMENTATION

- Segmentation (splitting word in many tokens) is widely used at translating into morphological rich languages (Oflazer and El-Kahlout, 2007; Ramanathanet al., 2008).
- Segmentation is carried out as a preprocessing, and then a usual SMT system is built. After translation a postprocessing stage is necessary to generate the final Basque words.
- We then use n-best list reranking in order to incorporate a word-level language model (since the language model used at decoding is based on the segmented text).
- We have tried different segmentation options, all of them based on the same morphological analysis obtained by Eustagger (Aduriz and Díaz de Ilarraza, 2003). Taking into account the morphemes obtained on the Eusatagger analysis, we have grouped them in a different way, leading to four different segmentations.



Analysis obtained by Eustagger for 'aukeratzerakoan' (*/at the election time/*) word and the different segmentation inferred from it:

Analysis:	aukeratu<adi><sin>+<adize>+<ala>+<gel>+<ine>				
Eustagger Segm.:	aukeratu<adi><sin>	+<adize>	+<ala>	+<gel>	+<ine>
Automatic Grouping:	aukeratu<adi><sin>	+<adize>+<ala>	+<gel>	+<ine>	
Hand-defined Grouping:	aukeratu<adi><sin>+<adize>	+<ala>+<gel>+<ine>			
One-Suffix Segm.:	aukeratu<adi><sin>	+<adize>+<ala>+<gel>+<ine>			

EUSTAGGER SEGM.

In this segmentation we strictly based on the lexicon of Eustagger and we have created a separate token for each morpheme recognizeg by the analyzer.

HAND GROUPING

We have hand deffined an intermediate morpheme grouping based on an analysis of the alignment errors. Defining a small amount of rules to morpheme grouping.

AUTOMATIC GROUPING

We have used Mutual Information metric to measure the statistical dependence between two morphemes. And we have grouped those that are more dependent than a threshold.

ONE-SUFFIX SEGM.

Since Eustagger lexicon is too fine-grained, we have put together all suffixes linked to a lemma into one suffix. So, at splitting one word we have generate at most three tokens (prefixes, lemma. suffixes).

EVALUATION

- Consumer corpus: a collection of 1036 Spanish articles of the Consumer Eroski magazine and their translation into Basque. The corpus is randomly divided in three subsets (training, devel, and test)

Some statistics of the corpus:

		Sentences	words	morphemes	word-vocabulary	morph-vocab.
training	Spanish	58,202	1,284,089	-	46,636	-
	Basque		1,010,545	1,699,988	87,763	35,316
devel	Spanish	1,456	32,740	-	7,074	-
	Basque		25,778	43,434	9,030	5.367
test	Spanish	1,446	31,002	-	6,838	-
	Basque		24,372	41,080	8,695	5.170

- Four evaluation metrics are reported: BLEU, NIST, WER and PER.
- The system which uses the Eustagger segmentation does not outperforms the baseline.
- The rest Morpheme-Based systems obtain significantly better results.
- According to BLEU the best results are obtained by the Hand morpheme-grouping

	BLEU	NIST	WER	PER
Baseline	10.78	4.52	80.46	61.34
MorphemeBased-Eustagger	10.52	4.55	79.18	61.03
MorphemeBased-AutoGrupping	11.24	4.66	79.15	60.42
MorphemeBased-HandGrouping	**11.36**	4.69	78.92	60.23
MorphemeBased-OneSuffix	11.24	**4.74**	**78.07**	**59.35**

BLEU, NIST, WER and PER evaluation metrics

- There is a correlation between the amount of tokens generated at segmentation and the evaluation results.
- Both unsegmented and fully segmented text (which have the bigger token difference regarding the Spanish) obtaines the worst results.
- The intermediate segmentations gets better results as their token amount is getting closer to the Spanish token amount.

	Running tokens	Vocabulary size	BLEU
Original words	1,010,545	87,763	10.78
Hand grouping	1,546,304	40,288	**11.36**
One-Suffix segm.	1,558,927	36,122	11.24
Automatic grouping	1,580,551	35,549	11.24
Eustagger segm.	1,699,988	35,316	10.52

Correlation between token amount on train corpus and BLEU metric

CONCLUSIONS

- We proved that the quality of the translation varies significantly according to the segmentation used, even when the same analysis is used to generate all segmentations
- Surprisingly, the criteria based on considering each morpheme as a separate token obtains worse results than the baseline. But, the other segmentations tried at this work outperform the baseline.
- The best results are obtained by a hand defined segmentation where some morphemes are grouped. This segmentation was defined after analyzing of the word alignment errors.
- Results obtained by the automatic morpheme grouping and the OneSuffix segmentation are good enough to consider them for SMT of other language pairs.