

# Improving Word Alignment Training through Morpho-syntactic Analysis

**Deepa Gupta and Marcello Federico**

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica  
38050 Povo (Trento) - Italy  
<gupta, federico>@itc.it

# Investigation

- Handling of data sparseness and scarceness through word transformation into:
  - Lemma form
  - Lemma plus morpho-syntax form
  - Stem form

# EPPS Training Data

<b>Training data (large data track)</b>		
<b>Description</b>	<b>Spanish</b>	<b>English</b>
Sentences	1,281,427	
Running words	36,578,514	34,918,192
Vocabulary	138,734	95,119
Singletons	47,708	34,135
Aver. sent. len.	28.5 (max. 100)	27.2 (max. 100)
<b>Training data (small data track)</b>		
<b>Description</b>	<b>Spanish (1%)</b>	<b>English (1%)</b>
Sentences	13,360	
Running words	385,198	366,055
Vocabulary	20,995	15,010
Singletons	9951	6453
Aver. sent. len.	28.8	27.4

# Test Data Statistics

<b>Translation test data</b>					
Lang.	sent.	Running word.	Vocab.	Avg. len.	Max. len.
Spanish	840	24,454	3839	29.1	161
<b>Alignment test data</b>					
Spanish	400	124,94	2976	31.2	99
English		117,90	2537	29.5	90

## Tool for Lemmas and Morpho-Syntax

- FreeLing Tagger: Spanish and English
  - For instance, Spanish input sentence is:  
"¿ hay alguna observación ?"

Original word	Lemma/base form	POS tag+ morpho-attributes
¿	¿	Fia
hay	haber	VAIP3S0
alguna	alguno	DI0FS0
observación	observación	NCFS000
?	?	Fit

## Tool for Lemmas and Morpho-Syntax

- For instance, English input sentence is:

Input sentence: are there any comments ?		
Original word	Lemma/base form	POS tag+ morpho-attributes
are	be	VBP
there	there	NN
any	any	DT
comments	comment	NNS
?	?	Fit

## Tool for Stemming

- Snowball stemmer: Spanish and English
- Example:

Spanish sentence

Input sentence: ¿ hay alguna observación ?

stemmed output: ¿ hay algun observ ?

English Sentence

Input sentence: are there any comments ?

Stemmed output: ar there ani comment ?

## Lemma + Morpho-Syntax Information

- Reduced morpho-syntax information
  - All inflected forms of the Spanish language are not relevant for translating Spanish text into English text.
    - For example, the adjective "*bonitd*" (beautiful/pretty)→  
*bonita, bonitas, bonito, bonitos*.
  - For each POS tags, we counted which additional morphological attributes do not effect the translation
  - Methodology: evaluate entropy



## Irrelevant/Relevant POS Morpho-Syntax Attributes for Spanish

POS	Irrelevant POS morpho-attributes	Relevant POS morpho-attributes
Verb	type (principal, auxiliary)	mode, time, person, number, gender
Noun	type (common, proper, etc.), gender, case	number (singular, plural, invariable)
Determine	type (demonstrative, possessive etc.), person, possessor	gender, number
Pronoun	person, possessor, politeness	type, gender, number, case

- No relevant morpho-attributes for: adjective, preposition
- No morpho-attributes at all for: adverb, conjunction, interjection

## EPPS Corpus Statistics after Using Lemma, Lemma Plus Morpho-Syntax and Stems

Training Data (large data track)								
Lang.	words		lemmas		lemmas + morph-syntax		stems	
	vocab.	sing.	vocab.	sing.	vocab.	sing.	vocab.	sing.
Spanish	138,734	47,708	77,960	30,172	131,669	46,273	78,749	30,108
English	95,119	34,135	81,947	31,599	-	-	69,268	27,126
Training Data (small data track 1%)								
Spanish	20,995	9951	11,823	4923	19,393	9157	9818	3825
English	15,010	6453	11,776	5092	-	-	9955	4140

# Word-Alignment Training

- Training on different word transformations:
  - Original words (baseline)
  - Lemmas
  - Lemma + reduced morpho-syntax Spanish and English original word
  - Stems

## Word-Alignment Results for Large Data

	Eng→Spa			Spa→Eng			Union		
	$R_s$	$P_p$	AER	$R_s$	$P_p$	AER	$R_s$	$P_p$	AER
<b>Baseline</b>	72.28	92.20	18.57	72.53	89.98	19.38	78.15	85.85	17.98
<b>Lemmas</b>	71.84	93.17	18.50	72.80	91.70	18.54	76.73	87.90	17.82
<b>Spa lem + redPOS</b>	71.94	92.06	18.82	72.72	90.46	19.06	77.87	86.16	17.97
<b>Stems</b>	72.45	93.56	17.94	73.39	92.20	17.98	77.60	88.60	17.01

$P$ (recision),  $R$ (ecall),  $s$ (ure link),  $p$ (ossible link) and  $A$ (lignment)  $E$ (rror)  $R$ (ate)

# Word-Alignment Results for Small Data

	Eng→Spa			Spa→Eng			Union		
	R <sub>s</sub>	P <sub>p</sub>	AER	R <sub>s</sub>	P <sub>p</sub>	AER	R <sub>s</sub>	P <sub>p</sub>	AER
<b>Baseline</b>	63.07	78.25	29.82	62.92	76.40	30.75	73.02	67.85	29.81
<b>Lemmas</b>	65.58	82.07	26.82	67.63	82.65	25.39	73.74	73.39	26.45
<b>Spa lem + redPOS</b>	63.40	78.72	29.43	63.97	77.79	29.43	73.31	68.79	29.14
<b>Stems</b>	66.51	82.89	25.90	68.25	82.54	25.07	74.69	73.88	25.73

# Observations about Word-Alignment Training

- All considered word transformation methods improve word alignment quality, in both small and larger data tracks.
- Use of morpho-syntactic information gives bigger improvement in case of data scarceness.
- While stemming gives the biggest reduction in alignment error rate in both data conditions.

## Translation Results on Large Data Track

- Experiment conditions for Spanish to English translation system
  - All translation models words→words
  - Language model: 34.9M running words
  - Evaluation: true case, with punctuation

Different word alignments	BLUE(%)	NIST	WER(%)	PER(%)
Baseline (words)	52.64	10.48	36.00	26.77
Lemmas	52.35	10.45	36.23	27.03
Lemma+red.POS Spa. and Org. Eng.	52.86	10.50	35.80	26.69
Stems	52.85	10.50	35.92	26.87

## Translation Results on Small Data Track

- Experiment conditions for Spanish to English translation system
  - All translation model words→words
  - Language model: 11.7K running words
  - Evaluation: true case, with punctuation

Different word alignments	BLUE(%)	NIST	WER(%)	PER(%)
Baseline (words)	40.60	9.117	43.74	32.77
Lemmas	40.81	9.164	43.85	32.61
Lemma+red.POS Spa. and Org. Eng.	40.86	9.164	43.71	32.41
Stems	40.84	9.178	43.91	32.52




## Translation Results on Small Data Track

- Experiment conditions for Spanish to English translation system
  - All translation models words→words
  - Language model: 34.9M running words
  - Evaluation: true case, with punctuation

Different word alignments	BLUE(%)	NIST	WER(%)	PER(%)
Baseline (words)	44.10	9.443	41.79	31.65
Lemmas	44.80	9.548	41.50	31.23
Lemma+red.POS Spa. and Org. Eng.	44.59	9.513	41.32	31.32
Stems	44.60	9.535	41.48	31.17

# Concluding Remarks

- **Marginal improvement** of translation scores  
by the use of :
  - Word-alignment training with lemma, reduced tags, stems
  - Translation model on original words
- **In both data conditions**, best improvement is achieved by Spanish lemma plus reduced POS method.
- Significant reduction in alignment error rate  
  
Significant improvement in translation scores