IDENTIFICATION OF BILINGUAL SEGMENTS FOR TRANSLATION GENERATION

K. M. Kavitha[1,3], Luís Gomes[1,2] and José Gabriel P. Lopes[1,2]

[1] CITI (NOVA LINCS), Faculdade de Ciências e Tecnologia, FCT/UNL, 2829-516 CAPARICA, Portugal
[2] ISTRION BOX-Translation & Revision, Lda., Parkurbis, Covilhã, 6200-865, Portugal
[3] Department of Computer Applications, St. Joseph Engineering College Vamanjoor, Mangalore, 575 028, India

MOTIVATION

- The bilingual translation lexicon (EN-PT) acquired through a cycle of iterations over parallel text alignment, term translation extraction and validation forms the basis of the study.
- The lexicon thus acquired is **not complete** as it does not contain all possible translation pairs.
- The extraction techniques cannot handle what is not in a parallel corpora, unless we care about automatically learning and generalizing word and multi-word structures.
- Any particular extraction technique is not able to extract everything.

Example translations from the input lexicon:

Source Term (EN)	Target Term (PT)			
ensure	assegurar, asseguram assegurem	zelar	garantir, garantem	permitir, permitam permitem
ensures	assegura, assegure		garante, garanta	permite, permita
ensured	asseguradas, assegurados assegurado, assegurou asseguraram	X	garantidas garantidos, garantido	X
ensuring	assegurando	X	garantindo	permitindo

APPROACH

1. IDENTIFYING BILINGUAL SEGMENTS

✧ **Look for orthographically similar translations**

- Note the bilingual stems and pair of bilingual suffixes attached to it.

ensured ⇔ assegurou ensuring ⇔ assegurando

ensur ⇔ assegur ing ⇔ ando, ed ⇔ ou

✧ **Analyse the induced bilingual segments** with respect to their occurrences as bilingual stems and bilingual suffixes.

- Each candidate bilingual stem should attach to at least 2 unique morphological extensions (pair of bilingual suffixes).
- Each pair of bilingual suffixes should have been attached to at least 2 unique bilingual stems.

2. FILTERING

✧ **Gather all the bilingual suffixes** associated with each bilingual stem identified in step 1.

('ensur', 'assegur'): ('e', 'em'), ('ing', 'ando'), ('ed', 'ou')

✧ **Eliminate redundant groups**

- Count unique translations in the lexicon that begin with each of the bilingual stems.
- Retain the bilingual stems that allow higher number of transformations.

✔ ('ensur', 'assegur'):
('e', 'ar'), ('ed', 'ado'), ('ed', 'ados'), ('ed', 'ada'), ('ed', 'adas'), ('ing', 'ando'), ('es', 'e'), ('es', 'a'), ('e', 'am'), ('e', 'em'), ('ed', 'aram'), ('ed', 'ou')

✘ ('ensure', 'assegur'):
('', 'ar'), ('d', 'ado'), ('d', 'ados'), ('d', 'ada'), ('d', 'adas'), ('s', 'e'), ('s', 'a'), ('', 'am'), ('', 'em'), ('d', 'aram')

3. CLUSTERING

✧ **Group the set of bilingual stems sharing same suffix transformations**

- We adopted the Partition Approach provided by Clustering Tool kit CLUTO.

Verb-e/ar Cluster:

('e', 'ar'), ('e', 'arem'), ('e', 'am'), ('e', 'em'), ('es', 'e'), ('es', 'a'), ('ed', 'ada'), ('ed', 'adas'), ('ed', 'ado'), ('ed', 'ados'), ('ed', 'aram'), ('ed', 'ou'), ('ing', 'ando'), ('ing', 'ar')

Example Bilingual Stems:

('toggl', 'comut'), ('argu', 'afirm')

Adjective-ent/ente Cluster:

('ent', 'ente'), ('ent', 'entes')

Example Bilingual Stems:

('bival', 'bival'), ('adjac', 'adjac'), ('coher', 'consist')

4. GENERATION

✧ **Suggesting new translations relies on the clusters of bilingual stems and bilingual suffixes identified in the learning phase.**

Generation statistics for training sets of different size:

Training Set	Unique Bilingual Stems	Unique Bilingual Suffixes	Generated Pairs	Correct Generations	Inorrect Generations
36K	6,644	224	4,279	3,862	306
210K	24,223	232	14,530	2,283/2,334	20/2,334

Highly frequent bilingual suffixes identified for EN-PT bilingual stems with different training sets:

36k Training Set		210k Training Set	
Suffix Pair	**Frequency**	**Suffix Pair**	**Frequency**
('', 'o')	4,644	('', 'o')	15,006
('', 'a')	2,866	('', 'a')	9,887
('e', 'o')	1,685	('', 'as')	5,840
('', 'os')	1,362	('', 'os')	5,697
('', 'as')	1,339	('ed', 'ado')	4,760
('e', 'a')	1,297	('ed', 'ados')	4,221
('ed', 'ado')	1,001	('ed', 'ada')	4,193
('ed', 'ada')	868	('e', 'o')	4,159
('ed', 'ados')	814	('ed', 'adas')	4,051

Co-occurrence frequency for correct translations in the parallel corpus:

Co-occurrence Frequency	# of generated bilingual pairs	Co-occurrence Frequency	# of generated bilingual pairs
9	45	4	148
8	62	3	207
7	64	2	324
6	80	1	489

RESULTS

✧ **90% of generated translations were correct when both the stem and suffix pairs in the bilingual pair to be analysed are known.**

Examples of inadequate/incorrect generated translations:

Accepted-	Rejected
languages ⇔ linguísticas	rights ⇔ adequados
instructor ⇔ instrutores	replaced ⇔ substituida / -idas / -idos / -ido
ambassador ⇔ embaixadores	several ⇔ vário
include ⇔ contam	wants ⇔ quere
emerged ⇔ resultados	electrical ⇔ electrica

Related Publications:

Gomes, L., Pereira Lopes, J.G.: *Parallel texts alignment*. In: New Trends in Artificial Intelligence, 14th Portuguese Conference in Artificial Intelligence, EPIA 2009, pp. 513–524 (2009).

Aires, J.,Pereira Lopes, J.G., Gomes, L.: *Phrase translation extraction from aligned parallel corpora using suffix arrays and related structures*. In: Progress in Artificial Intelligence, pp. 587–597 (2009).

[3] Snyder, B., Barzilay, R.: *Unsupervised multilingual learning for morphological segmentation*, pp. 737–745. ACL (2008).

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA