Bilingual Example Segmentation based on Markers Hypothesis

Alberto Simões, José João Almeida

Departamento de Informática, Universidade do Minho
Campus de Gualtar, 4710–057 Braga
{ambs,jj}@di.uminho.pt

Abstract

The Marker Hypothesis was first defined by Thomas Green in 1979. It is a psycho-linguistic hypothesis defining that there is a set of words in every language that marks boundaries of phrases in a sentence. While it remains a hypothesis because nobody has proved it, tests have shows that results are comparable to basic shallow parsers with higher efficiency.

The chunking algorithm based on the Marker Hypothesis is simple, fast and almost language independent. It depends on a list of closed-class words, that are already available for most languages. This makes it suitable for bilingual chunking (there is not the requirement for separate language shallow parsers).

This paper discusses the use of the Marker Hypothesis combined with Probabilistic Translation Dictionaries for example-based machine translation resources extraction from parallel corpora.

Index Terms: Marker Hypothesis, Probabilistic Translation Dictionaries, Translation Examples, Machine Translation

1. Introduction

Machine Translation (MT) and Computer Assisted Translation (CAT) use previously translated documents, for example parallel corpora aligned at the sentence level or the usual CAT translation memories. Unfortunately not all systems are able to adapt bilingual big sentence pairs to new sentences that require translation. This lack of re-usability is the motivation for Example-Based Machine Translation, a MT approach that segments bilingual sentence pairs into smaller segments with higher re-usability. These segments we call *translation examples*.

There are different articles on translation examples extraction and generalization [1]. Sentence segmentation is generally undertaken with language parsers or directly with generalization approaches [2, 3].

There is other work [4] using the Markers Hypothesis [5] for this segmentation, but it is not dealing with the examples alignment or with Iberian languages.

The presented document uses Probabilistic Translation Dictionaries (PTD) [6] together with the Marker Hypothesis to segment translation units into smaller aligned chunks (translation examples).

2. Probabilistic Translation Dictionaries

One of the most important resources for MT is translation dictionaries. They are indispensable, as they establish relationships between the language atoms: words. Unfortunately, freely available translation dictionaries have small coverage and for minority languages, are quite rare. It is crucial to have an automated method for the extraction of word relationships.

Simões and Almeida [6] explain how a probabilistic *word alignment* algorithm can be used for the automatic extraction of probabilistic translation dictionaries. This process relies on sentence-aligned parallel corpora.

The algorithm is language independent and therefore can be applied to any language pair. Experiments were executed using diverse languages, which included Portuguese, English, French, German, Greek, Hebrew and Latin [7]. The algorithm is based on word co-occurrences and its analysis with statistical methods. The result is a probabilistic dictionary which associate words on two languages.

These dictionaries map words from a source language to a set of associated words (probable translations) in the target language. Given that the alignment matrix is not symmetric, the process extracts two dictionaries: from source to target language and vice-versa.

The formal specification for one probabilistic translation dictionary (PTD) can be defined as:

$$w_{\mathcal{A}} \mapsto (occs\,(w_{\mathcal{A}}) \times w_{\mathcal{B}} \mapsto \mathcal{P}(\mathcal{T}(w_{\mathcal{A}}) = w_{\mathcal{B}}))$$

Figure 1 shows two entries from the English:Portuguese dictionary extracted from the EuroParl[8] corpus. Note that these dictionaries include the number of occurrences of the word on the source corpus, and a probability measure for each possible translation.

$$europe \rightharpoonup 42583 \times \begin{cases} \textit{europa} & 94.7\% \\ \textit{europeus} & 3.4\% \\ \textit{europeu} & 0.8\% \\ \textit{europeia} & 0.1\% \end{cases}$$

$$stupid \rightharpoonup 180 \times \begin{cases} \textit{estúpido} & 47.6\% \\ \textit{estúpida} & 11.0\% \\ \textit{estúpidos} & 7.4\% \\ \textit{avisada} & 5.6\% \\ \textit{direita} & 5.6\% \end{cases}$$

Figure 1: *Probabilistic Translation Dictionary examples.*

Regarding these dictionaries it should be noted that, although we use the term translation dictionaries, not all word relationships on the dictionary are real translations. This is mainly explained by the translation freedom, multi-word terms and a variety of linguistic phenomena.

Notwithstanding the probabilistic nature of these dictionaries, there is work on bootstrapping conventional translation dictionaries using probabilistic translation dictionaries [9] and on the connection between dictionaries quality and corpora genre and languages [10].

3. The Marker Hypothesis

The Marker Hypothesis was first defined by Thomas Green [5]. It is a psycho-linguistic hypothesis stating that there is a set of words in every language that marks boundaries of phrases in a sentence.

English	***Portuguese***
on	em; sobre; em cima de; de; relativa
once	desde que; uma vez que; se
only	todavia; mas; contudo
onto	para; para cima de; em direcção a
other	outro; outra; outras; outros
our	nosso; nossa; nossos; nossas
ours	o nosso; a nossa; os nossos; as nossas
owing to	devido a: por consequência de; por causa de
own	próprio; ser proprietário
past	por; para além disso; fora de
pending	durante; até
per	por; através de; por meio de; devido a acção de
plus	mais; a acrescentar a; a adicionar a
round	em torno de; à volta de
sort of	espécie de; género de; tipo de; de certo modo
since	desde; desde que; depois que
some	algum; alguns; alguma; algumas
subject to	sujeito a
such	este; esse; aquele; isto; aquilo
supposing	supondo; se; no caso de; dada a hipótese de
than	de; que; do que; que não
that	aquele; aquela; aquilo; esse; essa; isso; ...
the	o; a; os; as

Table 1: Markers list excerpt.

The algorithm uses a set of marker words (these are closed-class words, like articles, conjunctions, pronouns, prepositions, numerals and some adverbs) and search for them in the sentence to find phrases boundaries.

To illustrate the algorithm consider the following simple sentence:

John spent all day playing with his friends.

The markers present on this sentence are the words "*all*", "*with*" and "*his*". These words are marked in the sentence:

John spent <u>all</u> day playing <u>with his</u> friends.

The extracted segments start with one or more marker word (or at the beginning of the sentence) and end right before the next set of markers (or at the end of the sentence). This sentence would be therefore split on three segments:

John spent / <u>all</u> day playing / <u>with his</u> friends

For our experiments we obtained an English list of marker words from MaTrEx [4] project, where the Marker Hypothesis is also being used.

The Portuguese list was created based on the English version and enriched after the analysis of some experiment results. Table 1 shows an extract of these lists.

To help the reader to evaluate the kind of segment extracted using this algorithm, tables 2 and 3 show the most common

Occur.	Marker	Remaining segment
34 137	da	comissão
17 277	do	conselho
16 891	da	união europeia
11 379	em	matéria
9 880	de	trabalho
9 850	da	união
9 479	no	sentido
8 465	da	europa
8 454	da	UE
8 004	do	parlamento

Table 2: *Most occurring segments in the Portuguese language (from a total of 3 070 398 segments).*

Occur.	Marker	Remaining segment
13 566	and	gentlemen
11 466	the	commission
11 079	in	order
9 182	to	make
8 712	to	be
8 356	to	do
7 992	of the	european union
7 941	of the	committee
7 814	to	say
7 574	with	regard

Table 3: *Most occurring segments in the English language (from a total of 3 103 797 segments).*

segments in EuroParl [11] version 2 for the Portuguese and English languages. Note that these results were obtained processing both sides of the parallel corpora in an independent way.

Some other tests were performed to analyze the more productive markers, as can be seen in table 4. This information is useful to tune the segment alignment algorithm.

4. Marker Hypothesis on Translation Units

If we consider a translation units (for instance, the example above and its translation), and perform segmentation based on the Marker Hypothesis, the obtained result is:

John spent / <u>all</u> day playing / <u>with his</u> friends

<u>O</u> João passou / <u>todo o</u> dia / <u>a</u> jogar / <u>com os seus</u> amigos

As can be seen, the number of segments is not the same in different languages. This means that an alignment methodology is needed. A basic approach would be the use of the well known sentence alignment algorithm [12], but this method uses just sentence (or segment) length information. As these segments have similar lengths this algorithm is not the best approach.

Given the availability of Probabilistic Translation Dictionaries that include relationship information between words in the two languages, it is possible to perform a better alignment task.

For the segments alignment it is created a matrix where each column represents a segment in the source language and each row represents a segment in the target language. Each cell is filled with the probability of the smaller segment (being it in the source or target language) has its translation in the bigger segment (algorithm presented in figure 2). Cells with higher values are selected as good alignment points and the translation

Portuguese		English	
815815	de	541197	to
557697	,	471332	the
468409	a	440903	of
352064	da	400417	,
297634	do	370161	and
232629	e	252298	of the
197922	que	214191	in
196801	o	152164	a
178537	em	131225	in the
156299	dos	112446	for
[...]		105992	that
35394	para a	92180	on
33079	que o	91033	to the
32213	de um	78264	we
31539	nos	70578	on the
31492	muito	67805	this
30805	às	65092	that the
$>$ 234 000 diff. markers		$>$ 198 000 diff. markers	

Table 4: *More productive markers.*

examples are extracted. This is shown in table 5. Note that this example is not typical, but shown here for explanation purposes only.

	this decision shall take effect	on 16 september 1999
a presente decisão produz efeitos	**23.18**	5.86
em 16	0.00	**76.41**
de setembro	0.00	**85.60**
de 1999	0.00	**84.10**

Table 5: *Alignment Matrix.*

As usual on statistical methods, the extracted examples are then sorted and counted. This number of occurrence is a statistical indicator of the alignment quality. Other translation measures can be used to rank the extracted segments.

5. Results analysis

From a total of 1 507 225 different translation examples extracted (an occurrence average of 1.6654) with alignment of one to one segment, table 6 presents the 15 most occurring ones.

From these 15 examples just two are not really correct. The first one occurs because the closing parenthesis should be considered a special marker, because it is related with the segment that appears before (unlike the other markers). The second bad example results from the fact that "is" is considered a marker in the English list, while its translation is not in the Portuguese list (all forms of the verb "haver") and that the original English list does not include "there" as a marker (although it should be).

Tables 7 and 8 show examples of one to two and two to one alignments. The stars mark the segment pairs that we evaluate as problematic. Most of these pairs are quite near translations with just one or two extra words.

As the difference on number of segments raises the alignment quality lowers. This fact is not directly related to the used method but with the translation style.

Data: Consider $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$ are two segments in language $\mathcal{A}$ and $\mathcal{B}$, with $length(s_{\mathcal{A}}) < length(s_{\mathcal{B}})$ and dic is a probabilistic translation dictionary.

function transProb(dic, $s_{\mathcal{A}}$, $s_{\mathcal{B}}$)

$sMarkers \leftarrow$ markers $(s_{\mathcal{A}})$
$tMarkers \leftarrow$ markers $(s_{\mathcal{B}})$
$markProb \leftarrow$ quality $(dic, sMarkers, tMarkers)$
$sText \leftarrow$ text $(s_{\mathcal{A}})$
$tText \leftarrow$ text $(s_{\mathcal{B}})$
$textProb \leftarrow$ quality $(dic, sText, tText)$
return $0.1 \times markersProb + 0.9 \times textProb$

end

function quality(Dic, Set_1, Set_2)

$sum \leftarrow 0$
for $w_{\mathcal{A}} \in Set_1$ **do**
  for $w_{\mathcal{B}} \in dom\,(\mathcal{T}_{dic}\,(w_{\mathcal{A}}))$ **do**
    if $w_{\mathcal{B}} \in Set_2$ **then**
      $sum \leftarrow sum + \mathcal{P}\,(w_{\mathcal{B}} \in \mathcal{T}_{dic}\,(w_{\mathcal{A}}))$
return $\frac{sum}{\text{size}(Set_1)}$

end

Figure 2: *Translation probability computation algorithm.*

6. Conclusions

The use of the Marker Hypothesis as a tool to segment natural text is easier than the use of complex shallow parser systems because it is easier to configure (easy to define what are or not markers) and it works almost "out of the box" with little adjustments. Also, it requires little knowledge about the specific language where it is being applied. This makes it versatile to be used on languages which have few resources.

The use of Probabilistic Translation Dictionaries (PTD) to perform segment alignment is quite efficient. Given that the PTD extraction is completely automatic consequently it is not a bottleneck for the full process.

The translation examples extracted are interesting (although they need an evaluation on a Machine Translation system). The alignment algorithm can be improved which means that translation examples quality can raise.

For close languages like Portuguese and Spanigh we expect to have better quality results. Unfortunately at the time of writing we did not have a list of markers for Spanish neither a fluent Spanish speaker.

Unfortunately these examples can not be used alone in an example-based machine translation system as the boundary friction problem [13] is not solved. After translated examples concatenation a concordancer should be used to uniform the sentence.

7. Acknowledgments

Part of this work was done in the scope of the Linguateca project, contract no. 339/1.3/C/NAC, jointly funded by the Portuguese government and the European Union.

Occur.	Portuguese	English
36 886	senhor presidente	mr president
8 633	senhora presidente	madam president
3 152	espero	I hope
2 930	gostaria	I would like
2 572	o debate	the debate
2 511	penso	I think
2 356	está encerrado	is closed
1 939	penso	I believe
1 932	muito obrigado	thank
1 854	em segundo lugar	secondly
1 809	gostaria	I should like
⋆ 1 638	) senhor presidente	mr president
⋆ 1 524	há	there
1 423	infelizmente	unfortunately
1 345	creio	I believe

Table 6: *Top 1 to 1 segment alignments.*

Occur.	Portuguese	English
253	caros colegas	ladies and gentlemen
147	senhores deputados	ladies and gentlemen
143	devo dizer	I have to say
142	lamento	I am sorry
105	congratulo-me	I am pleased
95	estou convencido	I am convinced
90	vamos agora proceder	we shall now proceed
⋆ 90	e senhores deputados	ladies and gentlemen
90	agradeço	I am grateful
79	e outros , em nome	and others , on behalf
76	refiro-me	I am referring
⋆ 72	muito obrigado	thank you very
71	congratulo-me	I am glad
70	passamos agora	we shall now proceed
66	não há dúvida	there is no doubt

Table 7: *Top occurring 1–2 segment alignments (from 360 065 different segments)*

Occur.	Portuguese / English
986	segue-se na ordem
	the next item
222	(a sessão é suspensa
	(the sitting was closed
169	senhor presidente em exercício
	mr president-in-office
148	da sessão de ontem
	of yesterday 's sitting
142	(o parlamento aprova a acta
	(the minutes were approved
⋆ 138	dos assuntos económicos e monetários
	and monetary affairs
113	a proposta da comissão
	the commission 's proposal
110	a proposta da comissão
	the commission proposal
106	período de perguntas
	question time
⋆ 101	, em nome , sobre a proposta
	, on behalf
100	dos direitos do homem
	of human rights
84	dos direitos da mulher
	on women 's rights
⋆ 72	da direita do hemiciclo
	from the right
67	por interrompida do parlamento europeu
	of the european parliament adjourned
67	é muito importante
	it is very important

Table 8: *Top occurring 2–1 segment alignments (from 542 671 different segments)*

8. References

[1] A. Way, "Translating with examples," in *Workshop on Example-Based Machine Translation*, M. Carl and A. Way, Eds., September 2001, pp. 66–80.

[2] R. D. Brown, "Adding linguistic knowledge to a lexical example-based translation system," in *Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, Chester, England, August 1999, pp. 22–32. [Online]. Available: http://www.cs.cmu.edu/ ralf/papers.html

[3] ——, "Automated generalization of translation examples," in *Eighteenth International Conference on Computational Linguistics (COLING-2000)*, 2000, pp. 125–131. [Online]. Available: http://www.cs.cmu.edu/ ralf/papers.html

[4] S. Armstrong, M. Flanagan, Y. Graham, D. Groves, B. Mellebeek, S. Morrissey, N. Stroppa, and A. Way, "MaTrEx: machine translation using examples," in *TC-STAR OpenLab Workshop on Speech Translation*, Trento, Italy, 2006.

[5] T. R. G. Green, "The necessity of syntax markers. two experiments with artificial languages." *Journal of Verbal Learning and Behaviour*, vol. 18, pp. 481–496, 1979.

[6] A. M. Simões and J. J. Almeida, "NATools – a statistical word aligner workbench," *Procesamiento del Lenguaje Natural*, vol. 31, pp. 217–224, September 2003. [Online]. Available: http://alfarrabio.di.uminho.pt/ albie/publications/sepln2003.pdf

[7] A. M. B. Simões, "Extracção de recursos de tradução com base em dicionários probabilísticos de tradução," Ph.D. dissertation, Escola de Engenharia, Universidade do Minho, Braga, 19 May 2008.

[8] P. Koehn, "EuroParl: a multilingual corpus for evaluation of machine translation," 2002, draft.

[9] X. G. Guinovart and E. S. Fontenla, "Técnicas para o desenvolvemento de dicionarios de tradución a partir de córpora aplicadas na xeración do Dicionario CLUVI Inglés-Galego," *Viceversa: Revista Galega de Traducción*, vol. 11, pp. 159–171, 2005.

[10] D. Santos and A. Simões, "Portuguese-English word alignment: some experiments," in *LREC 2008 — The 6th edition of the Language Resources and Evaluation Conference*. Marrakech: European Language Resources Association (ELRA)., 28–30, May 2008.

[11] P. Koehn, "EuroParl: A parallel corpus for statistical machine translation," in *Proceedings of MT-Summit*, 2005, pp. 79–86.

[12] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," in *Meeting of the Association for Computational Linguistics*, 1991, pp. 177–184.

[13] R. D. Brown, R. Hutchinson, P. N. Bennett, J. G. Carbonell, and P. Jansen, "Reducing boundary friction using translation-fragment overlap," in *MT Summit IX*, New Orleans, 2003.