FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

Translation Dictionaries Triangulation

Alberto Simões[1], *Xavier Gómez Guinovart*[2]

[1]Universidade do Minho
[2]Grupo TALG, Universidade de Vigo
ambs@di.uminho.pt, xgg@uvigo.es

Abstract

Probabilistic Translation Dictionaries (PTD) are translation resources that can be obtained automatically from parallel corpora. Although this process is simple, it requires the existence of a parallel corpora for the involved languages.

Minoritized languages have a limited amount of available resources. For example, while they can have a few parallel corpora, the number of parallel language-pairs uses to be restricted.

We defend that if a minoritized language A has a parallel corpus with a language B, and language B has a parallel corpus with another language C, then we can obtain a helpful probabilistic translation dictionary between A and C.

In this document we will formalize the probabilistic translation dictionaries triangulation, perform some experiments making the triangulation between Galician, English and Italian, and conclude with an evaluation of the proposed approach.

Index Terms: probabilistic translation dictionaries, parallel corpora

1. Introduction

Translation between languages require as a basic resource a set of translation dictionaries, that is, a mapping from words or terms from one language to the other language. The main problem is that these resources are hard to create manually (time consuming, error prone, and hand-work intensive). Fortunately there are some methods [1, 2, 3, 4, 5] to analyze words used in a parallel corpus and extract automatically translation dictionaries.

These dictionaries are commonly named as Probabilistic Translation Dictionaries (PTD) as they are created in a statistic base. Nevertheless, they proven to be useful and have been used for different tasks:

- [6] describes a method to bootstrap a conventional translation dictionary from a PTD. A PTD was created and a try-and-error approach was used to define a translation probability threshold for filtering purposes. The filtered dictionary was used for manual validation.
- [7] uses PTDs for a similar task, bootstraping a machine translation dictionary. In this case the manual validation was not required.
- [8] uses PTDs as a mechanism to present parallel concordances and guessing translations of the searched terms, highlighting them when presenting the search result.
- [9, 10] presents methods to extract bilingual terminology using traslation patterns and PTDs for translations alignment.
- [11] also uses PTDs as a mechanism to align chunks of text when creating translation examples.
- [12] uses PTDs for cross-language information retrieval.

While useful, PTDs can only be obtained automatically if we have access to a parallel corpora in the required languages. That is, when computing a probabilistic translation dictionary between languages A and B, we need access to a parallel corpora between languages A and B. Unfortunately not all the language pairs in the world have available parallel corpora.

Consider, for instance, the Galician language. While we can find fairly easily parallel corpora with English, French, Portuguese or Spanish, it is not as easy to find parallel corpora with other languages like Arabic, Italian, Dutch or Danish.

Nevertheless, there are some parallel corpora from English, Portuguese or Spanish to these other more "exotic" languages. The question we want to answer is if it is possible to use two pair of transitive corpora, say Galician–English and English–Italian, to compute a Galician–Italian PTD. The choice of English as pivot language in our experiment is motivated by the high number of parallel corpora avalaible for English with very different languages. On the other hand, we feel that if it is possible to go from a Romance to a Germanic language, and then go back to a Romance language, then the same approach would hopefully retrieve even better results if we could find a pivot language in the same language family of source and target language.

For the better understanding of the concept of Probabilistic Translation Dictionaries, section 2 shows a simple example and discusses their structure details. Section 3 will formalize the triangulation (or composition) approach, and includes a full composition example. In section 4 we present some experiments on composing a Galician–English dictionary with an English–Italian dictionary, and evaluate the obtained results. Finally, we conclude on section 5 with some comments and future work.

2. Probabilistic Translation Dictionaries

Probabilistic Translation Dictionaries (PTD) are extracted automatically from (sentence-aligned) parallel corpora. The process is completely automatic and already proven to scale for big corpora. It results in a pair of dictionaries: one mapping words from the corpus source-language to its target-language, and another mapping words from the corpus target-language to its source-language[1].

Each dictionary maps words from a source-language S to a set of possible translations on a target-language T. Each possible translation have an associated probability measure:

$$\mathcal{T}\,(\text{codificada}) = \begin{cases} \text{codified} & 62.83\% \\ \text{uncoded} & 13.16\% \\ \text{coded} & 6.47\% \\ \dots \end{cases}$$

[1]Refer to [4] for further discussion about why these methods compute a pair of dictionaries instead of just one.

Together with this information we keep track, for each word on the source-language, of the number of their occurrence in the corpus. Given that a parallel corpus alignment produces a pair of PTD it is possible to query the number of occurrences for any word (being it from the source or target language), and to compute the total number of tokens for each corpus.

We will define formally a PTD (one of the two extracted dictionaries) from a language A to a language B as:

$$\begin{array}{lcl} PTD & = & langA \hookrightarrow info \\ info & = & count : int \quad \times \\ & & trans : trans \\ trans & = & langB \hookrightarrow prob \\ langA & = & term \\ langB & = & term \end{array}$$

The extraction tools usually discard some translations during PTD extraction, given computer memory limitations, keeping track of the best k translations and discarding translations with probabilities bellow a specified threshold. In the context of this article, we used NATools [5] and its default configuration values: the algorithm computes a maximum of 8 translations per word (the more probable), and rejects translations below a probability threshold.

3. PTD Triangulation

As described earlier, a PTD maps words into some information. Therefore, we can consider that a PTD behaves just like a function, receiving a word, and returning the probable translations structure for this word.

Given two dictionaries, D_1 and D_2, which map respectively words from language A to language B and words form language B to language C, it is possible to apply some kind of function composition between the two dictionaries, creating a dictionary $D = D_1 \circ D_2$, which maps words from language A to language C. Check [13] for an alternative approach of dictionaries composition.

As PTD do not map words into words, but words into some information structure, some decisions should be made so that this composition can be performed:

- **What occurrence should have each word on the target dictionary?**
  At the moment, our decision is to use the original occurrence count for that word, from the first dictionary. Another option could be to discard the value, or to multiply it by some factor related to the occurrence of the pivot word (the word being used for the composition).

  In the future we plan to perform experiments on using the a factor based on the multiplication of occurrences from both source languages (on both dictionaries).

- **What probability should be associated to each possible translation?**
  Although we can argue that the events of translating from A to B, and from B to C are not independent, we decided to just multiply the translation probabilities from both dictionaries, as defined below in the composition formalization.

  One problem of this approach is that the obtained probabilities are smaller than the ones we would usually obtain with a direct extraction, given the probability multiplication. With this in mind, after the composition we perform the dictionary totalization: sum up all the translation probabilities, consider this total to be 100% and recompute each word translation probability.

For a better understanding consider the following diagram. It presents some of the possible translations for the Galician word "afluencia" in English, together with their translation probability in the Galician-English PTD. For each English translation, we queried the English–Italian PTD, and added the more probable Italian translations (and their translation probability). The last column, in bold, presents the probability for the Italian word to be a correct translation of the original Galician word.

afluencia	influx	18.6%	afflusso	48.9%	**= 9.1%**
			flusso	12.7%	**= 2.4%**
			flussi	4.7%	**= 0.9%**
	flow	12.9%	flusso	46.9%	**= 6.0%**
			flussi	9.9%	**= 1.3%**
			gravi	1.7%	**= 0.2%**
	inflow	6.1%	sfogo	24.2%	**= 1.5%**
			afflusso	16.8%	**= 1.0%**
			ascritto	14.7%	**= 0.9%**
	flood	5.9%	inondazioni	5.6%	**= 0.3%**
			flusso	4.4%	**= 0.3%**
			alluvione	2.8%	**= 0.2%**
	flows	4.7%	flussi	72.3%	**= 3.4%**
			flusso	1.6%	**= 0.1%**
			ondate	1.5%	**= 0.1%**

As the Italian translations for each of the English word might repeat, these values should be summed up. So, the final version of the triangulation task would result in:

afluencia	afflusso	10.08%
	flusso	08.73%
	flussi	05.51%
	sfogo	1.46%
	ascritto	0.89%
	inondazioni	0.33%
	gravi	0.22%
	alluvione	0.16%
	ondate	0.07%

For the sake of completeness, we present formalization of the composition operator in mathematical notation.

$$\text{compose}: PTD \times PTD \longrightarrow PTD$$
$$\text{compose}(A, B) \stackrel{\text{def}}{=} \left(\begin{array}{c} w \\ \text{composeI}\,(A\,(w)\,, B) \end{array} \right)_{w \in \text{dom}(A)}$$

$$\text{composeI}: info \times PTD \longrightarrow info$$
$$\text{composeI}(A, B) \stackrel{\text{def}}{=} info\,(count(A), \text{composeT}\,(trans\,(A)\,, B))$$

$$\text{composeT}: trans \times PTD \longrightarrow trans$$
$$\text{composeT}(A, B) \stackrel{\text{def}}{=} \textbf{let}\ D = \text{dom}\,(A)\ \textbf{in}\ \left(\begin{array}{c} w \\ \text{p}\,(A, B, t, w) \end{array} \right)_{t \in D, w \in \text{dom}(trans(B(t)))}$$

$$\text{p}: trans \times PTD \times str \times str \longrightarrow double$$
$$\text{p}(A, B, t, w) \stackrel{\text{def}}{=} \textbf{let}\ tB = trans\,(B\,(t))\ \textbf{in}\ A\,(t) \times tB(w)$$

4. Triangulation Evaluation

For our experiments we used two Galician–English parallel corpora, TECTRA and UNESCO, [14] which are part of the CLUVI Parallel Corpus (`http://sli.uvigo.es/CLUVI/`), and the English–Italian pair from EuroParl v5 [15] parallel corpora. Table 1 summarizes extracted dictionary sizes.

Corpus	CLUVI		EuroParl v5	
Lang. pairs[2]	GL–EN	EN–GL	EN–IT	IT–EN
Types	100 740	69 861	118 001	170 159
T. per word[3]	4.61	6.12	5.94	5.15
Average prob.[4]	56%	49%	52%	58%
Average occs.[5]	18	28	403	288

Table 1: Statistics for the DPTs used in the experiment.

Table 2 summarizes the composed dictionaries (after probabilities totalization), for both language directions, and the composed dictionaries after a drastic filtering process.

The filtering used the following heuristics:

- the dictionary were totalized (table 2 values for the simple composition were taken after the totalization process);
- Then, were removed:
  - non-word entries (symbols, numbers, etc.);
  - entries with words occurring less than 5 times;
  - translations with probabilities bellow 20%;
  - translations $t \in \mathcal{T}_1(w)$ unless $\exists w \in \mathcal{T}_2(t)$;

Dictionary	**Simple Comp.**		**Filtered Comp.**	
Languages	GL–IT	IT–GL	GL–IT	IT–GL
Types	88 211	149 521	4 511	4 559
T. per word	31.2	47.82	1.1	1.1
Average prob.	34%	32%	50%	50%
Average occs.	21	323	168	4 746

Table 2: Statistics for the triangulated dictionary, before and after the drastic filtering process.

The resulting dictionaries are quite small, but the process of enlarging them is simple: just loosen the limits. Nevertheless, these limits should be defined accordingly with the final dictionary application. For simple automated tasks, like cross language information retrieval, there are no big losses on precision using weaker dictionaries. In the other hand, if preparing a dictionary for automatic or human translation, we might prefer fewer words and higher translation quality. Table 3 summarizes a looser approach, using the previous heuristics, but removing only entries with less than 2 occurrences (so, ignoring words occurring just once), and removing translations with probabilities bellow 10%.

[2]Language pairs with Galician (GL), English (EN) and Italian (IT).

[3]Average number of translations per dictionary entry. NATools limit this value to 8, justifying the number of translations on the original dictionaries. During composition that limit does not exist.

[4]Average probability for best entry translations. Higher values mean better translation confidence (and lower number of possible translations per entry).

[5]Average number of word occurrences in the source corpus.

Dictionary	**Simple Comp.**		**Filtered Comp.**	
Languages	GL–IT	IT–GL	GL–IT	IT–GL
Types	88 211	149 521	10 559	10 781
T. per word	31.2	47.82	1.4	1.4
Average prob.	34%	32%	39%	38%
Average occs.	21	323	97	2 628

Table 3: Statistics for the triangulated dictionary, before and after the looser filtering process.

These two changes make the average probability for the first translation to drop (as we have much more entries with fewer translation probabilities), and the average number of occurrences to drop as well (as we are including a lot of new words with occurrences ranging from 2 to 4).

For the triangulation evaluation 100 entries were randomly selected from both filtered dictionaries. These entries were evaluated manually, with three distinct classes: good translations (ignoring inflection), bad translations, and doubtful translations (where the translation is almost good, but misses something, for example, incomplete translation of one word to two word translation). Table 4 show obtained results. The number of evaluated translation pairs is not 100, as some entries have more than one possible translation.

Filtering	**Drastic Filtering**		**Looser Filtering**	
Languages	GL–IT	IT–GL	GL–IT	IT–GL
Good	104	101	100	106
Doubtful	5	4	7	3
Bad	1	4	26	29
Precision	95%	93%	75%	73%

Table 4: Evaluation of composed dictionaries after drastic and looser filtering approaches.

Note that, while the looser filtering approach increased the number of bad and doubtful entries, the number of good entries is almost the same.

Follows some examples of entries obtained with this method. The bad translations were tagged with a star. While those translations are bad, their presence is easy to understand.

abandonados	abbandonato	25.3%
	abbandonate	11.5%
	abbandonata	10.5%
advertencias	avvertimenti	41.8%
	avvertenze	15.5%
impostos	imposte	23.3%
	fiscale⋆	21.2%
	tasse⋆	12.8%
xenital	mutilazioni⋆	20.8%
	genitali	15.8%

5. Conclusions

Scarcety of linguistic resources is one of the problems of minoritized languages. In this paper, we have suggested a solution for that problem in the field of bilingual dictionaries construction, using probabilistic translation dictionaries (PTD) extracted from transitive parallel corpora.

With that aim, we have analyzed the issues concerning to the construction of a Galician–Italian probabilistic dictionary.

As we have seen, PTDs are bilingual dictionaries extracted automatically from parallel corpora on a statistic base. Nevertheless, while we can expect easily to find parallel corpora from Galician to Portuguese, English or Spanish, it is not that easy (or possible) to find available parallel corpora from Galician to Italian.

Our claim is that it is possible to use some kind of transitivity to create translation dictionaries. So we have shown how a Galician-Italian PTD can be constructed without a Galician–Italian parallel corpus, by the combination or triangulation of two other PTDs: a Galician–English PTD and an English–Italian PTD extracted, respectively, from a GL–EN parallel corpus and an EN–IT one.

Next we have evaluated the performance of the triangulation and, as expected, the final combined dictionaries are not as good as the ones extracted directly from parallel corpora. Even so, we can conclude that the process is great for bootstrapping dictionaries when better data is not available.

At the moment we are working on the treatment of the extracted combined dictionaries to populate an Italian–Galician bilingual dictionary similar to the on-line corpus-based CLUVI English–Galician Dictionary developed at the University of Vigo [16] (`http://sli.uvigo.es/dicionario`) This work will include the automatic extraction of usage examples candidates, and the conversion of the filtered version (looser approach) of the PTD to the XML format being used by CLUVI dictionaries. The final product of this lexicographic work will require a process of human revision, during which better measures on the precision of the triangulation procedure will be calculated.

As an option to make the dictionary larger and better, we could use more than one composition (for instance, $GL \rightarrow EN \rightarrow IT$, $GL \rightarrow ES \rightarrow IT$ and $GL \rightarrow PT \rightarrow IT$, and other paths that could be found) to retrieve a set of dictionaries that can be compared and merged in an enhanced Italian–Galician PTD.

6. Acknowledgements

This work was partially funded by the project *Português em paralelo com seis línguas (Português, Español, Russian, Français, Italiano, Deutsch, English)* grant `PTDC/CLE-LLI/108948/2008` from *Fundação para a Ciência e a Tecnologia*; and by the project *Desenvolvemento e explotación de recursos integrados da lingua galega* grant `INCITE08PXIB302185PR` from *Consellería de Innovación e Industria, Xunta de Galicia*.

7. References

[1] P. Fung and K. Church, "Kvec: A new approach for aligning parallel texts," Kyoto, Japan, pp. 1096–1102, 1994.

[2] N. Varma, "Identifying word translations in parallel corpora using measures of association," Master's thesis, University of Minnesota, 2002.

[3] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[4] D. Hiemstra, "Using statistical methods to create a bilingual dictionary," Master's thesis, Department of Computer Science, University of Twente, August 1996.

[5] 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

[6] X. G. Guinovart and E. S. Fontenla, "Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos," *Procesamiento del Lenguaje Natural*, vol. 33, pp. 133–140, 2004. [Online]. Available: http://webs.uvigo.es/sli/arquivos/sepln04a.pdf

[7] H. M. Caseli, M. G. V. Nunes, and M. L. Forcada, "Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts," *Procesamiento del Lenguaje Natural*, September 2005.

[8] A. Simões and J. J. Almeida, "NatServer: a client-server architecture for building parallel corpora applications," *Procesamiento del Lenguaje Natural*, vol. 37, pp. 91–97, September 2006. [Online]. Available: http://alfarrabio.di.uminho.pt/~albie/publications/sepln06.pdf

[9] ——, "Bilingual terminology extraction based on translation patterns," *Procesamiento del Lenguaje Natural*, vol. 41, pp. 281–288, September 2008.

[10] X. G. Guinovart and A. Simões, "Terminology extraction from English-Portuguese and English-Galician parallel corpora based on probabilistic translation dictionaries and bilingual syntactic patterns," in *I Iberian SLTech 2009*, A. Teixeira, M. S. Dias, and D. Braga, Eds., Porto Salvo, Portugal, September, 3–4 2009, pp. 13–16. [Online]. Available: http://webs.uvigo.es/sli/arquivos/IberianSLT09.pdf

[11] A. Simões and J. J. Almeida, "Bilingual example segmentation based on Markers Hypothesis," in *I Iberian SLTech 2009*, A. Teixeira, M. S. Dias, and D. Braga, Eds., Porto Salvo, Portugal, September, 3–4 2009, pp. 95–98.

[12] W. Kraaij, "TNO at CLEF-2001: Comparing Translation Resources," *Lecture Notes in Computer Science*, pp. 78–93, 2002.

[13] ——, "Exploring transitive translation methods," in *Proceedings of DIR*, 2003.

[14] X. G. Guinovart, "A investigación en lexicografía e terminoloxía no Corpus Lingüístico da Universidade de Vigo (CLUVI) e no Corpus Técnico do Galego (CTG)," in *A lexicografía galega moderna. Recursos e perspectivas*, E. González, A. Santamarina, and X. Varela, Eds. Santiago de Compostela: Consello da Cultura Galega / Instituto da Lingua Galega, 2008, pp. 197–216. [Online]. Available: http://webs.uvigo.es/sli/arquivos/sli_ilg07.pdf

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

[16] X. G. Guinovart, E. D. Rodríguez, and A. Álvarez Lugrís, "Aplicacións da lexicografía bilingüe baseada en córpora na elaboración do Dicionario CLUVI inglés-galego," *Viceversa*, vol. 14, pp. 71–87, 2008. [Online]. Available: http://webs.uvigo.es/sli/arquivos/viceversa_clig2ed_2008.pdf