Building a Basque-Chinese Dictionary by Using English as Pivot

Xabier Saralegi, Iker Manterola, Iñaki San Vicente
Elhuyar R & D
Zelai haundi 3, Osinalde Industrialdea
20170 Usurbil (Basque Country), Spain
E-mail: x.saralegi@elhuyar.com, i.manterola@elhuyar.com, i.sanvicente@elhuyar.com

Abstract

Bilingual dictionaries are key resources in several fields such as translation, language learning or various NLP tasks. However, only major languages have such resources. Automatically built dictionaries by using pivot languages could be a useful resource in these circumstances. Pivot-based bilingual dictionary building is based on merging two bilingual dictionaries which share a common language (e.g. L_A-L_B, L_B-L_C) in order to create a dictionary for a new language pair (e.g L_A-L_C). This process may include wrong translations due to the polisemy of words. We built Basque-Chinese (Mandarin) dictionaries automatically from Basque-English and Chinese-English dictionaries. In order to prune wrong translations we used different methods adequate for less resourced languages. Inverse Consultation and Distributional Similarity methods were chosen because they just depend on easily available resources. Finally, we evaluated manually the quality of the built dictionaries and the adequacy of the methods. Both Inverse Consultation and Distributional Similarity provide good precision of translations but recall is seriously damaged. Distributional similarity prunes rare translations more accurately than other methods.

Keywords: Bilingual Lexicons, Comparable Corpora, Less-Resourced Languages

1. Introduction

Bilingual dictionaries[1] are a key resource in a multilingual society. Their straight application can be found in a range of activities such as translation, language learning, etc. or as a basic resource for NLP tasks. However, the availability of such resources varies depending on the pair of languages. That is why most dictionaries include big languages such as English, Spanish, Chinese, etc., whereas dictionaries for languages with fewer speakers are scarce or even non-existent. When they exist, they are often limited to a major language (e.g. English-Basque) or to sociologically related languages (e.g Spanish-Basque, French-Basque...). Therefore, there are not many dictionaries which include two non-major languages (e.g Basque-Turkish) or even minor languages combined with some major languages (e.g. Basque-Chinese, Basque-Russian, Basque-Arab...). Economic considerations or the lack of great demand are the reasons for this. Automatically built dictionaries could be a useful resource in these circumstances.

Pivot-based bilingual dictionary building is based on merging two bilingual dictionaries which share a common language (e.g. L_A-L_B, L_B-L_C) in order to create a dictionary for a new language pair (e.g L_A-L_C). However, this process may include wrong translations due to the polisemy of words. A pivot word can lead to wrong translations corresponding to senses not represented by the source word (See Figure 1). These senses can be completely different or related but with a narrower or wider meaning.

In this work we use the same methods (Inverse Consultation and Distributional Similarity) as in (Saralegi et al., 2011) for building a Basque-Chinese dictionary via English. These methods are focused on less resourced languages. They just depend on easily available resources such as dictionaries including one major language and comparable corpora. In addition, we provide a manual evaluation of the resulting dictionaries. In the automatic evaluation performed by Saralegi et al. (2011) it was observed that several correct pairs were marked as wrong because they were not included in the reference dictionary. A manual evaluation allows us to measure the precision of the resulting dictionary more accurately. Furthermore, there is no Basque-Chinese dictionary that could be used as a reference.

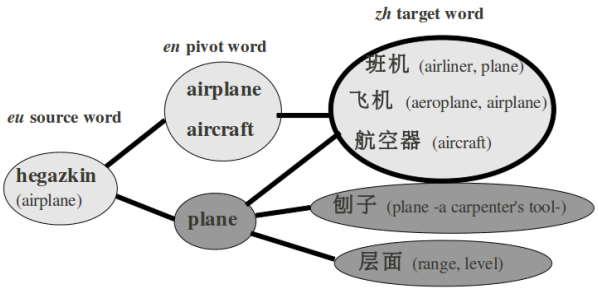

Figure 1: Ambiguity problem

This paper is structured as follows. The next section describes related works. Next the experimental setup is explained. After that, both methods used for pruning wrong pairs (Inverse Consultation and Distributional Similarity) are explained. The sixth section presents the evaluation of the dictionaries and the results are discussed. Finally, some conclusions are drawn.

2. Related work

In order to solve the ambiguity problem (See Figure 1.) some methods are proposed in the literature (Tanaka and

[1] In this paper we use a very flexible definition of dictionary. Using stricter lexicographic criteria it can be considered a list of bilingual equivalences.

Umemura, 1994; Shirai and Yamamoto, 2001; Bond et al., 2001; Kaji et al., 2008; Shezaf and Rappoport, 2010; Saralegi et al., 2011). Tanaka and Umemura (1994) propose the using of a number of pivot words between the source word and the translation candidate in order to measure the strength of the equivalence. They call this method Inverse Consultation (IC). All translation pairs failing to achieve a minimum number of pivot words are pruned. Other authors propose the pruning of wrong pairs in accordance with translation probability (Tsunakawa et al., 2008) or Distributional Similarity (Kaji et al., 2008; Shezaf and Rappoport, 2010). Distributional Similarity (DS) is adequate for less resourced languages because it does not require parallel corpora but rather comparable corpora, which are easier to obtain. Saralegi et al. (2011) analyze the requirements and properties of both IC and DS approaches and propose their combination to obtain better results. Other methods use definitions (Sjöbergh, 2005) or additional resources such as Wiktionary (Mausam et. al., 2009) for pruning wrong pairs. Unfortunately, such resources are often not available for minor languages.

3. Experimental setup

The experimental setup comprises several resources. A Basque-English[2] dictionary $D_{eu\rightarrow en}$ ={(w_{eu},w_{en})} and a Chinese-English[3] one $D_{zh\rightarrow en}$ ={(w_{zh},w_{en})} are used as sources to create the noisy Basque-Chinese dictionary candidate $D_{eu\rightarrow zh}$={(w_{eu},w_{zh})} (See Table 1) by through transitivity. Thus, $D_{eu\rightarrow zh}$ includes all pairs obtained by combining the equivalent-pairs of $D_{eu\rightarrow en}$ and $D_{zh\rightarrow en}$ which share at least one English equivalent:

$$D_{eu\rightarrow zh} = \{(w_{eu}, w_{zh}) : (w_{eu}, w'_{en}) \in D_{eu\rightarrow en} \wedge (w_{zh}, w_{en}) \in D_{zh\rightarrow en} \wedge w'_{en} = w_{en}\}$$

We can appreciate in table 1 that the average number of translations for each headword is significantly higher in the noisy dictionary $D_{eu\rightarrow zh}$ than in the initial dictionaries $D_{eu\rightarrow en}$ and $D_{zh\rightarrow en}$. Only 1,953 headwords include a single translation. These pairs are correct in a very high percentage because they are usually monosemous words. The rest of them (11,591) tend to include wrong translations.

Dictionary	#entries	#pairs	ambiguity[4]
$D_{eu\rightarrow en}$	17,699	42,994	2.43
$D_{zh\rightarrow en}$	37,313	63,899	1.71
(noisy) $D_{eu\rightarrow zh}$	13,544	182,089	13.44

Table 1: Dictionaries

The DS method needs comparable corpora in order to compute cross-lingual distributional similarity. We built comparable corpora for Basque and Chinese taking news from newspapers for the same period of time (2008-2011). *Berria*[5] was used for Basque and the *Beijing Daily*[6] for Chinese. The Basque corpus C_{eu} was lemmatized with *Eustagger* (Ezeiza et al., 1998). The Chinese corpus C_{zh} was segmented with the *Stanford Chinese Segmenter* (Tseng et al., 2005). Both C_{zh} and the Chinese part of $D_{zh\rightarrow en}$ were in simplified script.

Corpus	#doc	#token
C_{zh}	86K	55M
C_{eu}	150K	37M

Table 2: Comparable corpora

4. Inverse Consultation

Inverse Consultation (Tanaka and Umemura, 1994) is applied over the $D_{eu\rightarrow zh}$ noisy dictionary. The Inverse Consultation (IC) method is based on estimating the equivalent strength by measuring the number of pivot words between the source word and the translation candidate. We can see in Figure 1. that wrong translations (*"刨子"*, *"层面"*) are linked to the source word (*"hegazkina"*) just by a single pivot word (*"plane"*). By contrast, correct translations ("班机", "飞机", "航空器") are linked to source words by two pivot words (*"airplane"*, *"aircraft"*). The hypothesis behind IC is that that if there is more than one pivot word these words are lexical variants of the same senses. So source and target words share the same sense. Tanaka and Umemura (1994) established a minimum of 2 pivot words for guaranteeing correct translations. Thus, a pair candidate (s_{eu},t_{zh}) included in $D_{eu\rightarrow zh}$ is correct when:

$$\left|\{x_{en} : (s_{eu}, p_{en}) \in D_{eu\rightarrow en} \wedge (t_{zh}, p_{en}) \in D_{zh\rightarrow en} \wedge x_{en} = p_{en}\}\right| > 1$$

This method requires dictionaries including more than one lexical variant for each sense of equivalents in order to obtain a good performance.

5. Distributional Similarity

Distributional similarity (DS) is calculated from the bilingual comparable corpora (C_{eu},C_{zh}). At first, all the words corresponding to source w_{eu} and target words w_{zh} included in noisy dictionary $D_{eu\rightarrow zh}$ are represented by vectors $c(w_{eu})$ and $c(w_{zh})$ that include context words from the corpora C_{eu} and C_{zh}. Context words are selected according to a distance window (±5 tokens). The vector includes for each context word the association degree with respect to the word represented by the vector. Association degree is measured by log-likelihood ratio (Dunning, 1993). In order to measure the similarity between words in different languages one vector is projected to the other's language. The noisy bilingual dictionary $D_{eu\rightarrow zh}$ is used (See table 1) to translate vectors from Basque to Chinese $tr(c(w_{eu}))$. We select the most frequent translation in the Chinese Corpus C_{zh} for ambiguous translations. Then, the cosine distance between $tr(c(w_{eu}))$ and $c(w_{zh})$ vectors is computed. Those which do not reach a threshold are removed. The threshold can be tuned in accordance with the desired metric and using single translation pairs as reference. We used two thresholds: a low or flexible one and a high or strict one (TOP3).

[2] Elhuyar Basque-English dictionary

[3] Mdbg Chinese-English dictionary (simplified alphabet)

[4] Average number of translations per headword

[5] http://www.berria.info/

[6] http://www.bjd.com.cn/

6. Evaluation

We built different pruned $D_{eu \rightarrow zh}$ dictionaries according the proposed methods. Pruned dictionaries were created by flexible ($DSF\_D_{eu \rightarrow zh}$) and strict DS ($DSS\_D_{eu \rightarrow zh}$), IC ($IC\_D_{eu \rightarrow zh}$) and also by combining those methods ($ICDS\_D_{eu \rightarrow zh}$). As mentioned above, single translation entries are generally correct, and that is why, they were included in all dictionaries. The baseline consisted of not applying any pruning method ($D_{eu \rightarrow zh}$).

Different aspects of the automatically built dictionaries can be evaluated. We focused on two aspects: recall of the dictionary in terms of included headwords and translations, and average recall and precision of translations per entry[7]. The first aspect gives an idea of the coverage of the dictionary both in terms of headwords and translations. However, it does not reveal much about the quality of the equivalent-pairs. The second aspect provides information regarding the quality of each entry. Average recall and precision of translations per headword were computed to measure this second aspect. Those target words that share a sense with the headword were regarded as correct translations. In addition, we also analyzed whether most probable translations were included. More strict lexicographic criteria are not taken into account.

6.1 Recall of headwords and translations

For measuring the recall of headwords and translations provided by each method, we used the source dictionaries $D_{eu \rightarrow en}$ and $D_{zh \rightarrow en}$ as reference. According to the results, IC offers poor recall compared with DS (See table 3 and 4). Recall improves when both methods are combined. This means that results of both methods are partially divergent.

Dictionary	#headwords	R
$D_{eu \rightarrow zh}$	13,544	**0.76**
$IC\_D_{eu \rightarrow zh}$	3,574	0.20
$DSF\_D_{eu \rightarrow zh}$	9,767	0.55
$DSS\_D_{eu \rightarrow zh}$	9,767	0.55
$ICDS\_D_{eu \rightarrow zh}$	10,124	0.57

Table 3: Recall of headwords

Dictionary	#translations	R
$D_{eu \rightarrow zh}$	26,929	**0.72**
$IC\_D_{eu \rightarrow zh}$	4,607	0.12
$DSF\_D_{eu \rightarrow zh}$	19,592	0.52
$DSS\_D_{eu \rightarrow zh}$	14,638	0.38
$ICDS\_D_{eu \rightarrow zh}$	20,102	0.54

Table 4: Recall of translations

6.2 Recall and precision of translations per headword

For measuring the average recall and precision of translations per headword a reference was prepared manually. As this work is very time-consuming, only a random sample (150 entries) of the candidate dictionary ($D_{eu \rightarrow zh}$) was prepared. Frequency of use of headwords (according to C_{eu}) was also taken into account when random selection was performed. It is better to deal effectively with frequent words and frequent translations than rare ones. The Basque Corpus C_{eu} was lemmatized and POS tagged in order to extract the frequency information of the lemmas. Three frequency intervals were established: low frequency, medium frequency, high frequency (See table 5). 50 headwords were taken from each interval. The proportion between single translation entries and several translation entries was also maintained when the sample was prepared. In order to fairly compare the performance of IC and DS methods, only the entries that can be treated by both methods were included.

entries	Low frequency $(0 \leq f \leq 20)$	Medium frequency $(20 < f < 250)$	High frequency $(f \geq 250)$	All $(f \geq 0)$
unambiguous	1,065	651	237	1,953
ambiguous	4,681	3,53	3,385	11,591
all	5,746	4,176	3,622	13,544

Table 5: Number of entries of $D_{eu \rightarrow zh}$ wrt frequency intervals

All translation candidates for the sample 150 headwords were analyzed in order to calculate the precision *P(e)* and recall *R(e)* for each headword *e*. 8 lexicographers and translators took part in the manual annotation. Altogether, the sample has 3,407 pairs. We split the sample in 8 pair-sets and duplicated them. Each set received two different judges. That way we had two judges for each pair. No one was speaker of Chinese. They were native in Basque and they have advanced knowledge of English. So we used Chinese-English dictionaries (Yellowbridge[8], nciku[9]) including English definitions and examples to judge the correctness of each pair. We established a four-category system to perform the annotation:

a) Wrong pair: Source and target words do not share any sense.
b) Correct pair: Source and target words share one or more senses.
c) Different POS: Source and target words include same senses but different POS.
d) Doubt:
    1. Source and target words, refer to similar senses, but with narrower or wider meanings.
    2. Not enough information in the dictionaries to judge.

To judge whether two words of different languages are equivalent is difficult. Sometimes senses are not completely equal, or a sense in one language can be broader than in another. So it is very difficult to establish clear criteria to establish a line between wrong and right translations. As a result, some evaluators were more flexible than others, and different judges were assigned to the same pairs (See table 6). In order to solve the

[7] "Entry" refers to the set comprised by a headword and all its corresponding translations.

[8] http://www.yellowbridge.com/chinese/chinese-dictionary.php

[9] http://www.nciku.com/

detected cases of disagreement, the same annotators discussed the judgment until an agreement was reached. Only the critical cases of disagreement (wrong-correct) were discussed. The final sample used as the gold standard comprised those pairs not including any doubtful judgments (See table 7).

A annotator/B annotator	*Wrong*	*Correct*	*Different POS*	*Doubt*
Wrong	**860**	612	147	300
Correct		**1184**	153	367
Different POS			**164**	44
Doubt				**75**

Table 6: Agreement level for annotation of pairs

judges	*Wrong*	*Correct*	*Different POS*
#pairs	1070	1377	169

Table 7: Agreement level for pairs after discussion

Different measures useful for different use-cases were calculated:

- ***AvgF*$_1$***:* Average *F-score*.
- ***AvgF*$_{0.5}$**: Average *F-score* where precision *P* is weighted higher for all entries. Useful for scenarios where precision is critical.
- ***AvgF*$_2$**: Average *F-score* where recall *R* is weighted higher for all entries. Useful for scenarios where recall is critical.

$$AvgF_\beta = \frac{1}{|D_{eu\rightarrow zh}|} \cdot \sum_{e \in D_{eu\rightarrow zh}} (1+\beta^2) \frac{P(e)R(e)}{(\beta^2 \cdot P(e)) + R(e)}$$

Dictionary	***AvgR***	***AvgP***	***AvgF*$_1$**	***AvgF*$_{0.5}$**	***AvgF*$_2$**
$D_{eu\rightarrow zh}$	**1.0**	0.70	**0.82**	**0.74**	**0.92**
$IC\_D_{eu\rightarrow zh}$	0.34	**0.93**	0.49	0.69	0.39
$DSF\_D_{eu\rightarrow zh}$	0.74	0.73	0.73	0.73	0.74
$DSS\_D_{eu\rightarrow zh}$	0.35	0.74	0.47	0.61	0.39
$ICDS\_D_{eu\rightarrow zh}$	0.75	0.73	0.74	0.73	0.74

Table 6: Results for different metrics

Dictionary	**Low frequency** $(0 \le f \le 20)$	**Medium frequency** $(20 < f < 250)$	**High frequency** $(f \ge 250)$
$D_{eu\rightarrow zh}$	0.81	0.81	0.80
$IC\_D_{eu\rightarrow zh}$	0.52	0.51	0.45
$DSF\_D_{eu\rightarrow zh}$	0.73	0.73	0.74
$DSS\_D_{eu\rightarrow zh}$	0.5	0.49	0.42
$ICDS\_D_{eu\rightarrow zh}$	0.73	0.74	0.75

Table 7: *AvgF*$_1$ scores depending on frequency of headword *f(e)*

The results show (See table 6) that, surprisingly, the baseline is very competitive in all scenarios. IC is only competitive in high precision required scenarios (*AvgP*). DSF offers a more robust performance but it only outperforms the baseline in *AvgP*. We surmise that this could be due to the manually built reference which includes rare translations as correct. This fact gives the baseline very high recall and precision scores difficult to surpass. Probably because of the same reason and in contradiction with results of (Saralegi et al., 2011) DSS does not provide a significantly better *AvgP* than DSF either. As for the combination of IC and DS, it provides a slight improvement on *AvgR*.

Dictionary	**N**	**V**	**Adj.**	**Adv.**
$D_{eu\rightarrow zh}$	0.86	0.69	0.84	0.81
$IC\_D_{eu\rightarrow zh}$	0.56	0.29	0.49	0.46
$DSF\_D_{eu\rightarrow zh}$	0.78	0.61	0.72	0.72
$DSS\_D_{eu\rightarrow zh}$	0.55	0.23	0.28	0.51
$ICDS\_D_{eu\rightarrow zh}$	0.78	0.61	0.75	0.73

Table 8: *AvgF*$_1$ scores depending on POS of headword

We also analyzed how the performance of each method varies depending on the frequency of use of the headword *f(e)* obtained from C_{zh}. The results (See table 7.) show that frequent headwords are the most difficult to treat in the case of IC and DSS. In any case, the performance of all methods is quite robust regarding the frequency of source words. Otherwise, the performance of the methods varies significantly depending on the POS of the source word. According to the results (See table 8.) all methods show the best performance when dealing with nouns. Worst performance is obtained when verbs are treated.

Following manual analysis we saw that most errors related to IC consist of translations which have different POS from source word's (e.g. "*laido*" (*n*) (*insult*) → 侮辱(*v*) (*to insult*)). DS also has this problem. In addition, it includes many hypernyms or hyponyms as correct translations because they have high context similarity scores (e.g. "*mamu*" (*fancy dress*) → 服装(*dress*)).

6.2.1 Are most used translations included?

The baseline provides a good performance in terms of recall of headwords and translations (See tables 3. and 4.) and also for the average precision and recall of translation of each entry (See table 6.). However $D_{eu\rightarrow zh}$ includes some entries which have many translations, although many of them are very rarely used ("*eraman*" for example provides 281 Chinese translations). There are around of 2,500 headwords including more than 20 translations. Many users (e.g., foreign language learners in initial stages) would appreciate dictionaries comprising only the most probable translations. The quality of a dictionary that includes rare translations but which does not have the most widely used ones would be questionable[10]. In that scenario DS can be more useful than the baseline because unlike the baseline it does rank the translations. For measuring to what extent each dictionary includes only the most probable translations we designed a variant of recall which we call strict recall R_s. For each headword only the best scored translations are considered. TOP3 are selected in case of DS ($DSS\_D_{eu\rightarrow zh}$). In the case of the baseline three translations are randomly selected ($R3D_{eu\rightarrow zh}$), because no ranking of translations is available.

As for the reference, we obtained most probable translations from a parallel corpora composed of Basque

[10] We understand most probable translations of a source word as the most used lexical variants of the most used sense of the source word

and Chinese editions of the Bible. Such parallel corpora have been used in some basic Machine Translation systems (Phillips, 2001). Even if it is a small corpus (31,102 segments or verses) and has a relatively small and restricted vocabulary for creating a dictionary, it allows us to obtain the most probable translations for a number of words. We identified the most probable translations for the 1,596 Basque words whose frequency in the Bible corpus is greater that 10 ($\{w_{eu}: f(w_{eu})>10\}$). We accepted as most probable translations for a source word w_{eu} the set comprised by all the translations that exceed a probability ratio of 0.8. This ratio is computed between the probability of the translation ($p(w_{zh}|w_{eu})$) and the maximum translation probability ($max_{xzh}(p(x_{zh}|w_{eu}))$):

$\{w_{zh}: p(w_{zh}|w_{eu})/max_{xzh}(p(x_{zh}|w_{eu})) > 0.8\}$

Results show (table 8) that DS (TOP3) is more effective to keep only most probable translations on the dictionaries. So although it offers a poor average recall of translations per entry (See table 6) it is useful for creating more precise dictionaries where coverage of the most probable translations is more critical.

Dictionary	$AvgR_s$
$R3D_{eu\rightarrow zh}$	0.28
$DSS\_D_{eu\rightarrow zh}$	0.49

Table 8: $AvgR_s$ scores

7. Conclusions

This paper presents Basque-Chinese dictionaries. They were automatically created by means of pivot techniques, using IC and DS methods for pruning wrong translations. The quality of those dictionaries was manually evaluated. The pruning methods are useful for building dictionaries where precision of translations is required. IC is the most appropriate method for that propose. However, it suffers from low recall for translations and headwords. DS offers a poorer precision but a better balance between precision and recall. Nevertheless, the best balance between precision and recall is achieved when pruning methods (the baseline) are not applied. However, if we are interested in including only most probable translations DS offers a better performance than the baseline.

8. Acknowledgements

This work has been partially founded by the Industry Department of the Basque Government under grants IE09-262 (Berbatek project) and SA-2010/00245 (Pibolex+ project).

9. References

Dunning, T. (1994) Accurate Methods for the Statistics of Surprise and Coincidence. In Computational Linguistics 19(1): 61-74. Cambrigde, Mass: The MIT Press.

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. COLING-ACL.

Kaji H., Tamamura,S., Erdenebat D. 2008. Automatic construction of a Japanese-Chinese dictionary via English. LREC.

Mausam, Soderland S., Etzioni O., Weld D., Skinner M., Bilmes J. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. ACL.

Paik K., Shirai S., Nakaiwa H. 2004. Automatic construction of a transfer dictionary considering directionality. MLR.

Phillips, J. D. 2001. The bible as a basis for machine translation. In Proceedings of PACLing 2001.

Saralegi X., Manterola I., San Vicente I. 2011. Analyzing Methods for Improving Precision of Pivot-Based Bilingual Dictionaries. EMNLP.

Shezaf D., and Rappoport A. 2010. Bilingual lexicon generation using non-aligned signatures. ACL.

Sjöbergh J. 2005. Creating a free digital Japanese-Swedish lexicon. PACLING.

Tanaka K., Umemura K. 1994. Construction of a bilingual dictionary intermediated by a third language. COLING.

Tseng H., Chang P., Andrew G., Jurafsky D., Manning C. 2005. A Conditional Random Field Word Segmenter. SIGHAN.

Tsunakawa T., Okazaki N., Tsujii J. 2008. Building bilingual lexicons using lexical translation probabilities via pivot languages. LREC.