Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources

Yves Scherrer, Benoît Sagot

Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources

Yves Scherrer
Alpage, INRIA &
Université Paris 7 Diderot, Paris
yves.scherrer@inria.fr

Benoît Sagot
Alpage, INRIA &
Université Paris 7 Diderot, Paris
benoit.sagot@inria.fr

Abstract

We introduce a generic approach for transferring part-of-speech annotations from a resourced language to a non-resourced but etymologically close language. We first infer a bilingual lexicon between the two languages with methods based on character similarity, frequency similarity and context similarity. We then assign part-of-speech tags to these bilingual lexicon entries and annotate the remaining words on the basis of suffix analogy. We evaluate our approach on five language pairs of the Iberic peninsula, reaching up to 95% of precision on the lexicon induction task and up to 85% of tagging accuracy.

1 Introduction

Natural language processing for regional languages faces a certain number of challenges. First, the amount of electronically available written texts is small. Second, these data are most often not annotated, and spelling may not be standardized. One possible solution to these limitations lies in the use of an etymologically closely related language with more resources. However, in most such configurations, parallel corpora are not available since the languages are mutually intelligible and demand for translation is low.

In this paper, we present a generic approach for the transfer of part-of-speech (POS) annotations from a resourced language (RL) towards an etymologically closely related non-resourced language (NRL), without using any bilingual (i.e., parallel) data. We rely on two hypotheses. First, on the lexical level, the two languages share a lot of cognates, i.e., word pairs that are formally similar and that are translations of each other. Second, on the structural level, we admit that the word order of both languages is similar, and that the set of POS tags is identical. Thus, we suppose that the POS tag of one word can be transferred to its translational equivalent in the other language.

The proposed approach consists of two main steps. In the first step (Section 4), we induce a translation lexicon from monolingual corpora. This step relies on several methods, including a character-based statistical machine translation model to infer cognate pairs, and 3-gram and 4-gram contexts to infer additional word pairs on the basis of their contextual similarity. This step yields a list of $\langle w_{\text{NRL}}, w_{\text{RL}} \rangle$ pairs. In the second step (Section 5), the RL lexicon entries are annotated with POS tags with the help of an existing resource, and these annotations are transferred onto the corresponding NRL lexicon entries. We complete the resulting tag dictionary with heuristics based on suffix analogy. This results in a list of $\langle w_{\text{NRL}}, t \rangle$ pairs, covering the whole NRL corpus. A more detailed overview of our approach is available in Figure 1.

We evaluate our methods on five language pairs of the Iberic peninsula, where Spanish and Portuguese play the role of RLs: Aragonese–Spanish, Asturian–Spanish, Catalan–Spanish, Galician–Spanish and Galician–Portuguese.

2 Related work

Koehn and Knight (2002) propose various methods for inferring translation lexicons using only monolingual data. They consider several clues, including the identity or formal similarity of words (i.e., borrowings and cognates), similarity of the contexts of occurrence, and similarity of the frequency of words. They evaluate their method on English–German noun pairs. Our work is partly inspired by this paper, but uses different combinations of clues as well as updated methods and algorithms, and extends the task to POS tagging. We shall now describe in more detail the three major types of clues used in the literature.

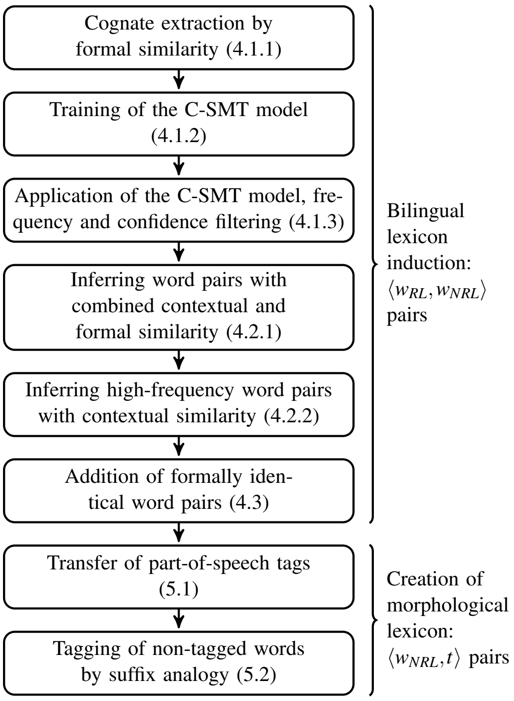
Cognate extraction by formal similarity (4.1.1)
Training of the C-SMT model (4.1.2)
Application of the C-SMT model, frequency and confidence filtering (4.1.3)
Inferring word pairs with combined contextual and formal similarity (4.2.1)
Inferring high-frequency word pairs with contextual similarity (4.2.2)
Addition of formally identical word pairs (4.3)
Transfer of part-of-speech tags (5.1)
Tagging of non-tagged words by suffix analogy (5.2)
Bilingual lexicon induction: ⟨w_RL, w_NRL⟩ pairs
Creation of morphological lexicon: ⟨w_NRL, t⟩ pairs

Figure 1: Flowchart of the proposed approach.

2.1 Cognate detection

Hauer and Kondrak (2011) define cognates as words of different languages that share a common linguistic origin. Two words form a cognate pair if they are (1) phonetically or graphemically similar, (2) semantically similar, and (3) if the phonetic or graphemic similarities are regular.

In closely related languages, cognates account for a large part of the lexicon. Mann and Yarowsky (2001) aim to detect cognate pairs in order to induce a translation lexicon. They evaluate different measures of phonetic or graphemic distance on this task. In particular, they distinguish static measures (independent of the language pair) from adaptive measures (adapted to the language pair by machine learning). Unsurprisingly, the authors observe better performances with the adaptive measures. However, they require a bilingual training corpus which we do not have at our disposal.

Kondrak and Dorr (2004) present a large number of language-independent distance measures in order to predict whether two drug names are confusable or not. Among the graphemic measures (they also propose measures operating on phonetic transcriptions), the BI-SIM algorithm (see Section 4.1.1) yields the best results. Inkpen et al. (2005) apply these measures to the task of cognate identification in related languages (English–French), and find that supervised classifiers do not perform better than language-independent methods with an accurately chosen threshold.

2.2 Character-based statistical machine translation

The principle underlying statistical machine translation (SMT) consists in learning alignments between pairs of words co-occurring in a parallel corpus. In phrase-based SMT, words may be grouped together to form so-called phrases (Koehn et al., 2003). Recently, a variant of this model has been proposed: character-based SMT, or henceforth C-SMT (Vilar et al., 2007; Tiedemann, 2009). In this paradigm, instead of aligning words (or word phrases) in a corpus consisting of sentences, one aligns characters (or segments of characters) in a corpus consisting of words. Of course, character alignments are well defined only for cognate pairs. Thus, it has been applied to translation between closely related languages (Vilar et al., 2007; Tiedemann, 2009) and to transliteration (Tiedemann and Nabende, 2009).

Whereas in the existing C-SMT literature training data is extracted from parallel corpora, we propose to create a (noisy) training corpus from monolingual corpora using cognate detection.

2.3 Context similarity

Exploiting context similarity is a promising approach for the induction of translation pairs from comparable corpora, whether the languages are closely related or not. The main idea (Fung, 1998; Rapp, 1999) is to extract word n-grams (or alternatively, bags of words) from both languages and induce word pairs that co-occur in the neighbourhood (context) of already known word pairs. For example, a French word appearing in the context of the word *école* is likely to be translated by an English word appearing in the context of the word *school*. This method requires a seed word lexicon (e.g., containing the pair $\langle \textit{école}, \textit{school} \rangle$), as well as large corpora in both languages in order to build sufficiently large similarity vectors.

Fišer and Ljubešić (2011) adapt this method to closely related languages: they build their seed lexicon with automatically extracted identical and similar words. Moreover, they take advantage of lemmatized and tagged corpora for both languages. Unfortunately, we lack annotated corpora for the non-resourced language — our goal is precisely to create such resources.

Context similarity methods have also been used in monolingual settings for lexical disambiguation (Bergsma et al., 2009) and for spelling correction (Xu et al., 2011): words that appear in similar contexts and are formally similar are likely to be alternative spellings of the same form. We pursue this idea in the area of closely related languages, where many word pairs not only are contextually similar, but formally as well.

2.4 Transfer of morphosyntactic annotations

The most straightforward idea for annotating a text from a non-resourced language consists in using a word-aligned parallel corpus, annotating the resourced side of it, and transferring the annotations to the aligned words in the other language. Yarowsky et al. (2001) successfully apply this approach to POS tagging, noun phrase chunking, named entity classification and even morphological analysis induction.

Another approach to this problem has been proposed by Feldman et al. (2006). They train a tagger on the resourced language and apply it to the non-resourced language, after some modifications to the tagging model. Such a tagger is bound to have a high OOV rate, and Feldman et al. (2006) propose two strategies to reduce it. First, they use a basic morphological analyzer for the non-resourced language to predict potential tags. Second, they extract a list of cognate pairs in order to transfer tags from one language to the other. While this approach looks promising, we chose to avoid the manual creation of a morphological analyzer, thus keeping our approach fully automatic.

3 Data

Our approach relies on three types of data:

1. A raw text of the NRL. From this text we extract word lists for cognate induction, frequency information by word-type as well as morphosyntactic contexts.

2. A raw text of the RL, from which we extract the same information.

3. A tag dictionary which associates RL words with their part-of-speech tags.

   We extract this dictionary from an annotated RL corpus; note however that tag dictionaries may be obtained from other sources, in which case no POS-annotated corpora are required at all by our approach.

Language	Sentences	Word tokens	Word types
Aragonese	335 091	5 478 092	215 809
Asturian	226 789	3 600 117	201 417
Galician	1 955 291	32 240 505	674 848
Catalan 200k	9 211	200 011	23 230
Catalan 500k	22 876	499 978	41 908
Catalan 1M	44 502	999 948	62 772
Catalan 10M	487 945	9 999 857	267 786
Catalan 50M	2 699 006	49 999 543	882 842
Catalan 140M	7 939 544	139 160 258	1 712 078
Spanish	23 381 287	431 884 456	3 451 532
Portuguese	12 611 706	197 515 193	2 252 337

Table 1: Wikipedia corpora

Language pair	Word types	Coverage
Aragonese–Spanish (AN–ES)	40 469	18.75%
Asturian–Spanish (AST–ES)	46 777	23.22%
Catalan–Spanish (CA–ES)	105 700	$\geq$ 6.17%
Galician–Spanish (GL–ES)	76 635	11.36%
Galician–Portuguese (GL–PT)	61 388	9.10%

Table 2: Size and coverage of the Apertium evaluation lexicons

The first two resources are used for the lexicon induction task, whereas the tag dictionary is required for the POS tagging task.

We test our approach on five language pairs: Aragonese–Spanish, Asturian–Spanish, Catalan–Spanish, Galician–Spanish and Galician–Portuguese, using raw text extracted from the respective Wikipedias. These language pairs vary widely in terms of available raw data and etymological distance, making them a good testing ground for our methods. Moreover, we use subsets of varying size of Catalan–Spanish to assess the impact of the data size (see Table 1).

We evaluate all five language pairs on the lexicon induction task on the basis of the dictionaries made available through the Apertium project (Forcada et al., 2011) (see Table 2).

The Spanish tag dictionary is extracted from the AnCora-ES corpus (Taulé et al., 2008).[1] It contains 42 part-of-speech tags and covers 40 148 words. The Portuguese tag dictionary is extracted from the CETEMPúblico corpus (Santos and Rocha, 2001).[2] It contains 117 part-of-speech tags (of which 48 are combinations of two tags) and covers 107 235 words.

[1] http://clic.ub.edu/corpus/ancora
We have slightly modified the AnCora corpus to split multi-word expressions and tag their components separately.

[2] http://www.linguateca.pt/CETEMPublico/

The Catalan–Spanish subsets are evaluated on the POS tagging task, using the AnCora-CA treebank as a gold standard. It is annotated according to the same guidelines as its Spanish counterpart.

4 Bilingual lexicon induction

In this section, we describe the different methods used for bilingual lexicon induction: the C-SMT method in Section 4.1, the n-gram context method in Section 4.2, and the addition of identical words in Section 4.3. Separate evaluations of the two former methods are presented in Sections 4.1.4 and 4.2.3 respectively.

4.1 Inferring cognate word pairs with character-based SMT

C-SMT models are generative models that translate words of the source language into their cognate equivalents in the target language. They are trained on a list of cognate word pairs, typically extracted form a word-aligned parallel corpus. Since we do not have bilingual data at our disposal, we propose to extract potential cognate pairs from two monolingual corpora (Section 4.1.1). Our hypothesis is that even with this noisy training data, the SMT models will learn useful generalizations. Section 4.1.2 describes the tools and parameters used for training the C-SMT model. Section 4.1.3 introduces two filters designed to further improve the precision of C-SMT.

For practical reasons, we infer the cognate pairs in the direction $w_{\text{NRL}} \rightarrow w_{\text{RL}}$, i.e., we consider the NRL as the source language and the RL as the target language. In particular, this allows us to match different w_{NRL} with the same w_{RL} and thus to take into account orthographic variation in the NRL. Such variation is less expected in the RL, which is assumed to have standardized spelling. Moreover, the classic SMT architecture puts the resource-intensive language model on the target language side, which is an additional argument in favour of the chosen translation direction.

4.1.1 Cognate extraction by formal similarity

We start by extracting word lists from the Wikipedia corpora. For the source language, we remove short words (< 5 characters) and hapaxes. For the target language, we remove short words and words with less than 1000 occurrences.[3]

The formal similarity between two words is computed with the BI-SIM measure (Kondrak and Dorr, 2004). BI-SIM is a measure of graphemic similarity which uses character bigrams as basic units. It does not support swap operations, and it is normalized by the length of the longer string. Thus, it captures a certain degree of context sensitivity, avoids crossing alignments and favours associations between words of similar length. This measure is completely generic and does not presuppose any knowledge of the etymological relationship between the two languages. In contrast, it is not very precise and yields highly ambiguous results. For example, the Catalan–Spanish word pairs $\langle \textit{activitat}, \textit{actividad} \rangle$ and $\langle \textit{activitat}, \textit{activista} \rangle$ yield the same BI-SIM value, even if only the former can be considered a cognate pair.

For each source word w_{NRL}, we keep the $\langle w_{\text{NRL}}, w_{\text{RL}} \rangle$ pair(s) that maximize(s) the BI-SIM value, but only if this value is above the (empirically chosen) threshold of 0.8. This threshold allows us to remove unlikely correspondences. When several w_{NRL} are associated with the same w_{RL}, we keep all of them. The resulting list of cognate pairs is then used as training corpus for the C-SMT model.

4.1.2 Training of the C-SMT model

Our C-SMT model relies on the standard pipeline consisting of GIZA++ (Och and Ney, 2003) for character alignment, IRSTLM (Federico et al., 2008) for language modelling, and Moses (Koehn et al., 2007) for phrase extraction and decoding. These tools may be configured in various ways; we have tested a large set of parameter configurations in preliminary experiments, but due to space restrictions, we just mention the parameter settings that we finally retained.

- We add special symbols to the beginning and the end of each word.
- We train a character 10-gram language model on the target language words. We removed words appearing less than 10 times in the corpus; each word is repeated as many times as it appears in the corpus.
- GIZA++ produces distinct alignments in both directions. Among the proposed heuristics, the *grow-diag-final* algorithm was the most efficient.

[3] This threshold has been introduced to reduce the complexity of comparing every source word with every target word. We have found that a lower threshold does not necessarily improve the results.

	Source	BI-SIM		C-SMT		Frequency filter		Confidence filter	
	words	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
AN–ES	92 393	34.52%	64.44%	100%	76.26%	100%	74.45%	94.04%	74.54%
AST–ES	77 517	43.02%	62.76%		75.02%		74.93%	89.11%	78.50%
GL–ES	280 828	23.68%	41.26%		69.57%		72.20%	90.66%	72.85%
GL–PT	280 828	14.93%	36.83%		48.89%		53.06%	89.90%	53.31%
CA–ES 200k	8 781	57.18%	68.03%	100%	70.42%	100%	69.08%	83.07%	77.37%
CA–ES 500k	16 456	52.83%	64.26%		70.92%		69.81%	82.18%	78.31%
CA–ES 1M	25 633	47.36%	60.39%		69.86%		69.29%	81.56%	78.01%
CA–ES 10M	111 232	27.34%	45.01%		62.93%		65.55%	88.05%	69.09%
CA–ES 50M	363 627	16.81%	37.61%		56.13%		62.19%	90.28%	63.18%
CA–ES 140M	750 287	11.81%	34.94%		51.52%		58.41%	89.30%	59.13%

Table 3: Evaluation of the cognate word induction steps. Recall refers to the percentage of source words for which the respective method yielded at least one target word. Precision refers to the percentage of correct pairs among the answered pairs whose source word appears in the evaluation lexicon.

- We have disallowed distortion (i.e., the possibility of changing the order of characters) to avoid learning crossing alignments, which we suppose very rare in the context of word correspondences between related languages.

- *Good Turing discounting* is used to adjust the weights of rare alignments.

- The different parameter weights of an SMT model are usually estimated through *Minimum Error Rate Training* on a development corpus. However, the tuned weights yielded worse results than the default weights, due to the large amount of noise in the training data. Thus, we kept the default weights.

4.1.3 Application of the C-SMT model and filtering

Once trained, the C-SMT model is used to generate a target word for each source word, using the same list of source words as for the creation of the training corpus. This is thus a completely unsupervised approach. Now, the C-SMT model may also generate RL words that occur less than 1000 times and that have been filtered out during training.

In line with the findings of Koehn and Knight (2002), preliminary experiments have shown that word pairs with large frequency differences are often wrong. For example, Catalan *coneguda* ‘known’ is associated with Spanish *conseguida* ‘reached’ instead of the more frequent (and correct) *conocida* ‘known’. Therefore, we generate a 50-best list of candidates with C-SMT and rerank them according to the frequency similarity between the source word and the target word. Frequency counts are extracted from the monolingual Wikipedia corpora. In the following, we refer to this as *frequency filtering*.

Moreover, the absolute C-SMT and frequency scores are a good indicator of the quality of the translation pair. These scores allow us thus to remove word pairs that are likely to be wrong, by adding a second filter. It eliminates all candidates whose combined score is less than 0.5 standard deviations below the mean of all combined scores. We call this *confidence filtering*.

4.1.4 Evaluation

Table 3 shows the results of the different lexicon induction steps described above.

Unsurprisingly, the training corpus extracted with the BI-SIM method is rather noisy, with precision values of less than 70%. Its recall values are low as well: target candidates were found only for 11% to 58% of the source words.

This picture changes impressively with the C-SMT model trained on these noisy data. Not only does it generate a candidate for each source word (recall of 100%), but the resulting precision values improve by about 10% absolute on average.

The frequency filter only seems to work reliably when the source language corpora are large enough (from the 10M Catalan subset onwards, and including Galician). The confidence filter improves precision values at the expense of recall; its relevance thus largely depends on the task at hand. Still, precision values are higher than 70% and recall values higher than 80% in most experiments.

Finally, one may note that, despite the filters, precision degrades drastically with very large source corpora (> 10M running words). This is likely to be caused by the addition of rare words, which are often named entities that do not follow the regular graphemic correspondences.

4.2 Inferring word pairs with contextual similarity

For several reasons, methods based on formal similarity alone are not always adequate: (1) even in closely related languages, not all word pairs are cognates; (2) high-frequency words are often related through irregular phonetic correspondences; (3) pairs of short words may just be too hard to predict on the basis of formal criteria alone; (4) formal similarity methods are prone to inducing false friends, i.e., words that are formally similar but are not translations of each other. For these types of words, we propose a different approach that relies on contextual similarity.

Suppose that our corpora contain the Catalan segment *diferència de càrrega elèctrica* and the Spanish segment *diferencia de carga eléctrica*. Suppose further that the C-SMT system has inferred the word pairs $\langle \textit{diferència}, \textit{diferencia} \rangle$ and $\langle \textit{elèctrica}, \textit{eléctrica} \rangle$. These word pairs allow us to match the two segments and to propose two new potential word pairs, $\langle \textit{de}, \textit{de} \rangle$ and $\langle \textit{càrrega}, \textit{carga} \rangle$. Other context pairs may then validate or invalidate these word pairs.

We use 3-gram context pairs of the type $\langle w_1 w_2 w_3, v_1 v_2 v_3 \rangle$, with already known word pairs $\langle w_1, v_1 \rangle$ and $\langle w_3, v_3 \rangle$, to infer the new word pair $\langle w_2, v_2 \rangle$. Likewise, we use 4-gram context pairs of the type $\langle w_1 w_2 w_3 w_4, v_1 v_2 v_3 v_4 \rangle$, with already known word pairs $\langle w_1, v_1 \rangle$ and $\langle w_4, v_4 \rangle$, to infer the new word pairs $\langle w_2, v_2 \rangle$ and $\langle w_3, v_3 \rangle$. We skip punctuation signs in the context construction.[4]

It is evident that word pairs inferred by matching contexts are extremely noisy. We therefore propose two filtering approaches: a filter based on both context frequency and formal similarity criteria for cognates and near-cognates (4.2.1), and a back-off filter based on frequency criteria alone for short high-frequency words (4.2.2).

4.2.1 Combined contextual and formal similarity

We filter the $\langle w, v \rangle$ word pairs obtained by context matching according to the following criteria:

- Word pairs inferred by one single context are not deemed reliable enough.
- We also remove word pairs with a relative string edit distance higher than 0.5.[5]
- For a given source word, we remove all contextually inferred target candidates in the lower half of their frequency distribution and in the lower half of their distance distribution. This allows us to focus on those candidates that are clearly more similar than their concurrents.

A lot of the retained word pairs have already been proposed by the C-SMT method. We found that 70%-80% of the contextually inferred word translations are identical to the C-SMT translations, whereas 10%-20% of word pairs are new, and the remaining 5%-15% concern source words which were translated differently with C-SMT. Among this last category, we mainly find different inflected forms of the same lemma, and different transliterations of the same named entity. However, the context approach also corrects some erroneous C-SMT pairs, such as Aragonese–Spanish $\langle \textit{charra}, \textit{carrera} \rangle$ 'talks/race', replacing it by the correct $\langle \textit{charra}, \textit{habla} \rangle$. Therefore, when merging the C-SMT word pairs and the context word pairs, we give precedence to the latter.

4.2.2 Removing the formal similarity criterion for high-frequency words

The combined filter unfortunately removes some high-frequency grammatical words that are either non-cognates (e.g. Catalan–Spanish $\langle \textit{amb}, \textit{con} \rangle$), or whose forms are too short to compute a meaningful distance value (e.g. $\langle i, y \rangle$ with a relative edit distance of 1.0). For these cases, we introduce a back-off filter that lacks the formal similarity criterion and focuses only on frequency cues.

Concretely, each source word that has not obtained a target candidate with the previous approach is assigned the target word with the high-

[4] It is also possible to use a 3-gram context in one language and a 4-gram context in the other one to infer word pairs of the type $\langle w_2, v_2 v_3 \rangle$ or $\langle w_2 w_3, v_2 \rangle$. Such patterns are useful if the two languages have different tokenization rules. For example, they have allowed us to obtain the Asturian–Spanish pairs $\langle \textit{a l'}, \textit{al} \rangle$ and $\langle \textit{polos}, \textit{por los} \rangle$. However, for the time being, we have not integrated such asymetric alignments in the evaluation framework and in the POS tagging pipeline.

[5] Since the contexts already constrain the potential word pairs, we chose to be more tolerant with the formal similarity criterion and explicitly use a lower threshold (0.5 instead of 0.8) and a simpler distance measure (string edit distance instead of BI-SIM) than above.

	Combined		High-frequency	
	Pairs	Precision	Pairs	Precision
AN–ES	3389	88.35%	35	34%
AST–ES	7549	92.56%	37	65%
GL–ES	22933	94.58%	91	67%
GL–PT	12518	87.04%	90	42%
CA–ES 200k	292	89.92%	7	40%
CA–ES 500k	915	94.77%	14	60%
CA–ES 1M	1676	94.80%	32	78%
CA–ES 10M	9065	94.03%	90	71%
CA–ES 50M	17014	92.96%	141	67%
CA–ES 140M	20514	91.87%	186	60%

Table 4: Evaluation of the word pairs induced by contextual similarity.

est number of common contexts, provided that this number is higher than 5.

Moreover, we have opted for a pigeonhole principle here: we disallow a target word to be matched with more than one source word. In our case, this prevents all pronouns to be assigned to the more frequent definite determiners.

This filter yields only a small number of word pairs, but they are of crucial importance since their token frequency is very high.

4.2.3 Evaluation

The performance of the context similarity approach is illustrated in Table 4.

The combined similarity method yields word pairs with very high precision. The number of induced word pairs grows according to the size of the corpus from which the contexts are extracted.

The high-frequency word approach works less well: the number of induced word pairs is very low, and translation precision falls drastically. While the quality of the word pairs induced with this approach may be insufficient for lexicon induction, we still deem it good enough for the POS tagging task. Indeed, the reliance on context similarity means that even if the induced word forms are wrong, they are still of the correct grammatical category. For example, the Galician–Spanish words $\langle boa, gran \rangle$ are not translations of each other but are both adjectives.

4.3 Addition of formally identical word pairs

Even after the application of the C-SMT and context lexicon induction methods, many words remain untranslated. (Remember that the recall figures of Table 3 refer to the number of source words used for this method, which excludes hapaxes and words with less than 5 characters.) For these words, we simply check whether they figure in identical form in the target language. This mainly allows us to add punctuation signs, but also abbreviations, numbers and proper nouns.

5 Creation of the morphological lexicon

In the preceding sections, we have described how we induce a bilingual lexicon from monolingual non-annotated texts. In this section, we use this lexicon to create a POS tag dictionary for the NRL, and use it to annotate texts.

5.1 Transfer of morphological annotations

The bilingual lexicon induced above contains $\langle w_{\mathrm{NRL}}, w_{\mathrm{RL}} \rangle$ pairs. Annotation transfer amounts to (1) loading an existing $\langle w_{\mathrm{RL}}, t \rangle$ tag dictionary for the resourced language, and (2) merging these two resources by transitivity in order to obtain $\langle w_{\mathrm{NRL}}, t \rangle$ pairs.

The tag dictionaries extracted from AnCora-ES (for Spanish) and from CETEMPúblico (for Portuguese) contain ambiguities, i.e. words that are assigned several part-of-speech tags depending on their syntactic function. For the time being, we do not deal with these ambiguities, but we rather associate each word unambiguously with its most frequent POS tag. With this simplification, merging the two dictionaries by transitivity is straightforward.

5.2 Adding morphological annotations by suffix analogy

At this point, there still remain untagged NRL words, either because no induced bilingual word pair contained it, or because the corresponding RL word was not found in the tag dictionary. In this case, we guess its tag by suffix analogy. We identify the longest suffix that is common to the non-annotated word and to at least one annotated word, and we transfer the POS tag of the annotated word to the non-annotated word. If several annotated words share the same suffix, we choose the most frequent POS tag.

5.3 Distribution of POS tag induction methods

Table 5 shows the percentage of word tokens and word types that have been tagged with the different tag induction methods. As already mentioned above, the C-SMT approach is mainly used for long low-frequency words that contain regular

	Tokens				Types			
	C-SMT	Context	Ident.	Suffix	C-SMT	Context	Ident.	Suffix
AN–ES	14.8%	49.1%	20.4%	15.7%	13.5%	1.6%	3.5%	81.4%
AST–ES	11.3%	54.2%	18.8%	15.7%	14.3%	3.6%	4.0%	78.2%
GL–ES	8.0%	59.1%	18.8%	14.1%	6.3%	2.6%	1.3%	89.8%
GL–PT	15.0%	55.2%	20.8%	9.0%	15.5%	2.2%	3.6%	78.7%
CA–ES 200k	17.7%	43.2%	20.5%	18.6%	14.8%	1.1%	16.9%	67.2%
CA–ES 500k	18.2%	47.9%	18.3%	15.6%	20.7%	3.0%	14.6%	61.7%
CA–ES 1M	17.4%	52.0%	17.2%	13.4%	24.4%	5.0%	12.9%	57.8%
CA–ES 10M	14.4%	62.3%	15.1%	8.3%	30.2%	16.3%	7.0%	46.5%
CA–ES 50M	14.2%	64.5%	14.1%	7.2%	29.3%	21.7%	5.4%	43.6%
CA–ES 140M	15.4%	63.5%	14.0%	7.1%	29.8%	21.8%	5.1%	43.4%

Table 5: Distribution of the origin of the induced POS tags, by word types and tokens.

phonetic correspondences. The contextual similarity methods are used for frequent words. The context methods account for more than half of the tokens, but for no more than 22% of the types. The *Identical* category mainly concerns punctuation signs, which again have high token frequencies. Finally, suffix analogy is used for the overwhelming majority of word types, but accounts for less than 20% of token frequencies.

The size of the source corpus impacts the distribution of the different tagging methods: the coverage of the context similarity methods increases, while the other methods are used less frequently.

5.4 Evaluation

Finally, we have evaluated the POS tagging accuracy of the Catalan–Spanish datasets, using AnCora-CA as a gold standard. The results range from 79.9% token accuracy with the smallest dataset up to 85.1% token accuracy with the largest one. All methods except suffix analogy yield accuracy rates higher than 70%. Given the difficulty of the task and the complete absence of annotated Catalan resources used in the process, these results can be considered satisfying.

As the corpus size increases, the highly accurate context similarity methods take over more and more words from C-SMT. For the remaining words, the C-SMT approach yields lower accuracy. However, this shift only has a small impact on the global accuracy rates, which seem to plateau at the 10M dataset. Adding more data above this threshold does not sensibly improve the results.

6 Conclusion

We have proposed a combination of several lexicon induction methods for closely related languages and have used the resulting lexicon to transfer part-of-speech annotations from a resourced language to a non-resourced one. Note that this task is more complex than the more traditional task of non-supervised part-of-speech tagging, for which a POS dictionary of the respective language is generally available. We have applied our methodology to five Romance language pairs of the Iberic peninsula and evaluated it on different subsets of our Catalan–Spanish data.

	C-SMT	Context	Ident.	Suffix	Total
200k	85.3%	91.7%	83.2%	43.3%	79.9%
500k	86.1%	91.1%	85.8%	45.2%	82.0%
1M	85.7%	90.4%	87.8%	46.9%	83.3%
10M	73.8%	90.8%	89.8%	51.2%	84.9%
50M	70.3%	90.0%	93.9%	52.3%	85.0%
140M	71.4%	90.1%	94.0%	53.6%	85.1%

Table 6: Token tagging accuracy on the Catalan–Spanish datasets.

Several aspects of this work may be improved. First, the assumed one-to-one correspondence between words and tags is clearly not satisfactory, and ambiguity should be introduced in a controlled way. This would also allow us to train a real POS tagger on the data, which could learn to disambiguate the words on the basis of the syntactic contexts and also tag unknown words more accurately than the suffix analogy method used here.

Second, we would like to replace the various threshold-based filters of the context similarity method by a more generic approach, possibly based on a classifier trained on the word pairs obtained with C-SMT. Unfortunately, first tests have resulted in insufficient recall.

Finally, we plan to validate our methodology on additional language pairs. We have started experimenting with Germanic and Slavic languages.

Acknowledgements

This work has been funded by the Labex EFL (ANR/CGI), operation LR2.2.

References

Shane Bergsma, Dekang Lin, and Randy Goebel. 2009. Web-scale n-gram models for lexical disambiguation. In *Proceedings of IJCAI 2009*, pages 1507–1512.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of Interspeech 2008*, Brisbane, Australie.

Anna Feldman, Jirka Hana, and Chris Brew. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 549–554.

Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'11)*, pages 125–131.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. *Machine Translation and the Information Soup*, pages 1–17.

Bradley Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 865–873, Chiang Mai, Thailand.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in French and English. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, pages 251–257.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition (SIGLEX 2002)*, pages 9–16, Philadelphia, PA.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT'03)*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), demonstration session*, Prague, République tchèque.

Grzegorz Kondrak and Bonnie Dorr. 2004. Identification of confusable drug names: A new approach and evaluation methodology. In *In Proceedings of COLING 2004*, pages 952–958.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, pages 151–158, Pittsburgh, PA, USA.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, Maryland, USA.

Diana Santos and Paulo Rocha. 2001. Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of ACL 2001*, pages 442–449.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of LREC 2008*.

Jörg Tiedemann and Peter Nabende. 2009. Translating transliterations. *International Journal of Computing and ICT Research*, 3(1):33–41. Special Issue of Selected Papers from the fifth international conference on computing and ICT Research (ICCIR 09), Kampala, Uganda.

Jörg Tiedemann. 2009. Character-based PSMT for closely related languages. In *Proceedings of the 13th Conference of the European Association for Machine Translation (EAMT 2009)*, pages 12 – 19, Barcelone.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, Prague, République tchèque.

Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, and Le Zhao. 2011. Exploiting syntactic and distributional information for spelling correction with web-scale n-gram models. In *Proceedings of EMNLP 2011*, pages 1291–1300, Edinburgh.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, San Diego, USA.