A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages

Yves Scherrer, Benoît Sagot

A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages

Yves Scherrer[1,2] **Benoît Sagot**[1]

1. Alpage, INRIA & Université Paris Diderot, Bâtiment Olympe de Gouges, 75013 Paris, France
2. LATL-CUI, Université de Genève, 7 route de Drize, 1227 Carouge, Switzerland
yves.scherrer@unige.ch, benoit.sagot@inria.fr

Abstract

In this paper, we describe our generic approach for transferring part-of-speech annotations from a resourced language towards an etymologically closely related non-resourced language, without using any bilingual (i.e., parallel) data. We first induce a translation lexicon from monolingual corpora, based on cognate detection followed by cross-lingual contextual similarity. Second, POS information is transferred from the resourced language along translation pairs to the non-resourced language and used for tagging the corpus. We evaluate our methods on three language families, consisting of five Romance languages, three Germanic languages and five Slavic languages. We obtain tagging accuracies of up to 91.6%.

Keywords: Part-of-speech tagging, lexicon induction, closely related languages

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 describe our latest experiments based on our generic approach for transferring 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 (Scherrer and Sagot, 2013).[1] This approach relies on two hypotheses. First, at 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, at the structural level, we admit that the word order of both languages is similar, and that the set of POS tags is identical. Under these hypotheses, the POS tag of one word can be transferred to its translational equivalent in the other language.

Our approach consists of two main steps. We first induce a translation lexicon from monolingual corpora, based on cognate detection followed by cross-lingual contextual similarity (Section 4). This step yields a list of $\langle w_{\mathrm{NRL}}, w_{\mathrm{RL}} \rangle$ translation pairs. Second, POS information from an existing RL lexicon is transferred along translation pairs (Section 5) and used for tagging the corpus; NRL words still lacking a POS are tagged based on suffix analogy. This architecture is summarized in Figure 1.

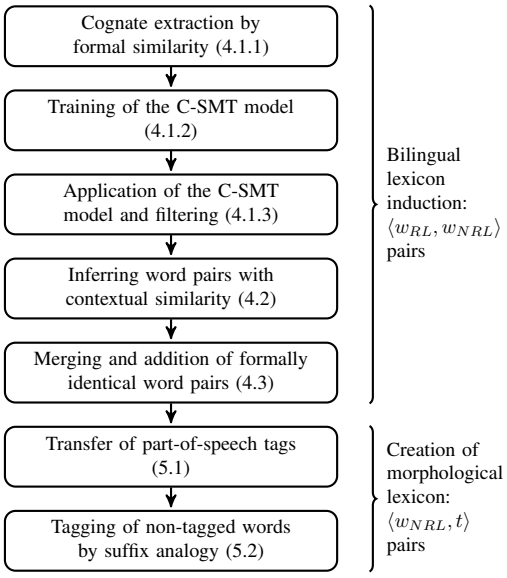

Figure 1: Flowchart of the proposed approach.

We evaluate our methods on three language sets:

- five Romance languages of the Iberic peninsula: Spanish and Portuguese play the role of RLs, whereas Aragonese, Asturian, Galician and Catalan[2] are considered NRLs;
- three Germanic languages, German being the RL whereas Palatine German (Pfälzisch) and Dutch are considered NRLs;
- five Slavic languages, with Czech and Polish as RLs and Slovak, Upper Sorbian, Lower Sorbian and Kashubian as NRLs.

[1] This task is different from unsupervised POS-tagging, where a morphosyntactic lexicon is usually admitted, and whose task is to disambiguate the annotation. In our work, we do not presuppose any ressource for the NRL except raw textual data.

[2] We introduce a small (500k words) and a large (140M words) Catalan corpus to study the impact of the corpus size (see Table 1).

Language	ISO	Raw corpora (Wikipedia) #sentences	#tokens	#types	Annotated corpora for gold lexicon extraction Name	#word types	#tags
Aragonese	AN	335,091	5,478,092	215,809	–		
Asturian	AST	226,789	3,600,117	201,417	–		
Galician	GL	1,955,291	32,240,505	674,848	–		
Catalan 500k	CA	22,876	499,978	41,908	–		
Catalan 140M	CA	7,939,544	139,160,258	1,712,078	–		
Portuguese	PT	12,611,706	197,515,193	2,252,337	CETEMPúblico[3]	107,235	117
Spanish	ES	23,381,287	431,884,456	3,451,532	AnCora-ES[4]	40,148	42
Dutch	NL	33,361	499,991	52,502	–		
Palatine German	PFL	28,149	318,926	51,038	–		
Standard German	DE	42,127,804	612,658,190	8,673,998	TIGER[5]	85,691	55
Kashubian	CSB	25,620	198,560	40,805	–		
Lower Sorbian	DSB	28,352	265,580	48,189	–		
Upper Sorbian	HSB	106,299	891,941	104,319	–		
Slovak	SK	2,555,779	30,114,232	1,091,474	–		
Czech	CS	6,642,402	85,579,006	1,934,787	PDT 2.5[6]	55,947	57
Polish	PL	16,639,594	206,372,541	3,264,129	NKJP[7]	132,664	29

Table 1: Corpora used in our experiments.

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 (borrowings and cognates), contextual similarity, and frequency similarity. 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.

Cognate pair extraction has been studied for example by Mann and Yarowsky (2001), using various phonetic and graphemic clues. Kondrak and Dorr (2004) compare several measures and introduce the BI-SIM graphemic measure, showing its relevance for assessing whether two drug names are confusable or not. Inkpen et al. (2005) apply these measures for cognate identification in related languages (English–French).

We use automatically extracted cognate pairs as a (noisy) training corpus for training a character-level SMT (henceforth C-SMT) system. In this paradigm, instead of aligning words (or word phrases) in a set of sentences, one aligns characters (or character sequences) in a corpus of words. C-SMT has been applied to translation between closely related languages (Vilar et al., 2007; Tiedemann, 2009), to transliteration (**?**) and to cognate generation (Beinborn et al., 2013).

Cross-lingual contextual similarity has also been used for inducing translation pairs from comparable corpora. The main idea (Fung, 1998; Rapp, 1999) is to extract word n-grams (or bags of words) from both languages and induce word pairs that co-occur in the neighbourhood (context) of already known word pairs. This method requires a seed word lexicon 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, but contrarily to us, they rely on a tagger and a lemmatizer for both languages. Context similarity has also been used in a monolingual setting, e.g., 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 extend this idea to cognates in closely related languages.

A lot of recent work has been dedicated to transferring POS annotations from one language to another using a word-aligned parallel corpus (Yarowsky et al., 2001; **?**; **?**). Since parallel data is not available for most language pairs covered in our experiments, these methods cannot be used without major adaptations.

Another approach consists in training a tagger on the resourced language and adapting the tagging model itself to the non-resourced language (Feldman et al., 2006). However, this approach is not entirely unsupervised since it requires a morphological analyzer for the NRL.

3. Data

Our approach relies on three types of data: a raw NRL text, a raw RL text and a tag dictionary which associates RL words with POS. In our experiments, we extract raw textual data from RL and NRL Wikipediæ and POS dictionaries from annotated RL corpora.[8] Raw corpora are used for the lexicon induction task, whereas the tag dictionary is required for the POS tagging task. A summary of the data we used in our experiments is given in Table 1.

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

[4] http://clic.ub.edu/corpus/ancora

[5] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

[6] http://ufal.mff.cuni.cz/pdt2.5/

[7] http://nkjp.pl

[8] 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.

4. Bilingual lexicon induction

In this section, we outline the methods used for both major lexicon induction steps, the one based on formal similarity (Section 4.1) and the one based on contextual similarity (Section 4.2).

4.1. Inferring cognate word pairs with character-level SMT

C-SMT models are generative models that translate words of the source language into their cognate equivalents in the target language, character by character. They are trained on a list of cognate word pairs whose characters are then aligned. The word pairs are typically extracted form a word-aligned parallel corpus, but 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 C-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 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, expected to be higher in the NRL than in the RL. We assume that the spelling of the RL is normalized.

4.1.1. Cognate extraction by formal similarity

We start by extracting word lists from the raw corpora, removing short words (words with less than 5 characters) as well as rare words (words accounting for the lowest 10% of the frequency distribution). The resulting lists contain around 10,000 to 25,000 words per language.

We then compute the above-mentioned BI-SIM score between each word of the NRL and each word of the RL. For each source word w_{NRL}, we keep the $\langle w_{\mathrm{NRL}}, w_{\mathrm{RL}} \rangle$ pair(s) that maximize(s) the BI-SIM value, provided it is above the (empirically chosen) threshold of 0.8.[9]

The BI-SIM measure is completely generic and does not presuppose any knowledge of the etymological relationship between the two languages, but is not very precise and therefore generates ambiguous results.[10] Hence, when a w_{NRL} is associated with several w_{RL}, we keep all of them.

4.1.2. Training of the C-SMT model

The word pairs extracted with the BI-SIM measure are then used to train a 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. In particular, the following settings yielded optimal results:

Beginning and end of word symbols We add special symbols at the beginning and at the end of each word.

Language model We have trained a character 10-gram language model on the words extracted from the RL corpus. We removed words appearing less than 10 times in the corpus; each word is repeated as many times as it appears in the corpus.

Alignment combinations GIZA++ produces distinct alignments in both directions. Among the proposed heuristics, the *grow-diag-final* proved the most efficient.

Distortion In the SMT terminology, distortion refers to the possibility of changing the order of elements. We disallow distortion altogether to avoid learning crossing alignments, which we suppose very rare between words of closely related languages.

Smoothing We use *Good Turing discounting* to adjust the weights of rare alignments.

Tuning The different parameter weights of an SMT model are usually estimated through *Minimum Error Rate Training* on a development corpus. Since we only have very noisy bilingual corpora, the resulting models performed less well than the ones with default weights. We chose to keep the latter.

4.1.3. Application of the C-SMT model and filtering

Once trained, we use the C-SMT model to generate a target NRL word for each source RL word, using the same RL word list as for the creation of the training corpus. In comparison with the word pairs extracted with BI-SIM, the C-SMT-generated word pairs obtain at the same time higher precision (the C-SMT model is sensitive to language-pair-specific regularities) and higher recall (it also translates words that were excluded by the abovementioned 10% frequency threshold or the 0.8 BI-SIM threshold).

In line with the findings of Koehn and Knight (2002), preliminary experiments have shown that word pairs with large frequency differences are often wrong. Therefore, we rerank the 50-best candidates obtained by the C-SMT model according to the frequency difference between the NRL and the RL word in the pair (*frequency filter*).

Moreover, the combined C-SMT and frequency scores serve as basis for an additional *confidence filter*. This confidence filter removes all word pairs whose combined score is below 0.5 standard deviations below the mean score.

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,

[9] This pairwise comparison is the most time-consuming step of the pipeline, which is why we filtered the word lists rather aggressively.

[10] For example, the Catalan–Spanish word pairs ⟨*activitat*, *actividad*⟩ and ⟨*activitat*, *activista*⟩ yield the same BI-SIM value, even if only the former can be considered a cognate pair.

3-gram	w_1	w_2		w_3
	$\|\|$	$\updownarrow$		$\|\|$
	v_1	v_2		v_3
Example AN-ES	tendrá	importans		conseqüencias
	$\|\|$	$\updownarrow$		$\|\|$
	tendrá	importantes		consecuencias
4-gram	w_1	w_2	w_3	w_4
	$\|\|$	$\updownarrow$	$\updownarrow$	$\|\|$
	v_1	v_2	v_3	v_4
Example CA-ES	diferència	de	càrrega	elèctrica
	$\|\|$	$\updownarrow$	$\updownarrow$	$\|\|$
	diferencia	de	carga	eléctrica

Figure 2: Illustration of the context extraction process. Word sequences on one line represent chunks of texts extracted from the corpus. The vertical equal signs link word pairs already inferred in the seed data, while the vertical double arrows link newly inferred word pairs.

we propose a different approach that relies on contextual similarity, using the word pairs obtained by the C-SMT system as seed data.

We extract 3-gram and 4-gram contexts from both languages and form context pairs whenever the first and the last word pairs figure in the seed data, allowing the word pair(s) in the center to be newly inferred. Figure 2 illustrates this process. We retain a newly inferred word pair if it has been seen in two or more distinct contexts.[11]

Word pairs inferred by matching contexts alone are 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 formal similarity value lower than 0.5.
- 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.

[11] 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 $\langle$*a l'*, *al*$\rangle$ and $\langle$*polos*, *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.

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 $\langle$*amb*, *con*$\rangle$), or whose forms are too short to compute a meaningful distance value (e.g. $\langle i, y \rangle$ with a formal similarity value of 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 highest 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.

4.3. Merging of the dictionaries and addition of formally identical word pairs

The word pairs induced through C-SMT and through context similarity overlap to a large extent: 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 $\langle$*charra*, *carrera*$\rangle$ 'talks/race', replacing it by the correct $\langle$*charra*, *habla*$\rangle$. Therefore, when merging the C-SMT word pairs and the context word pairs, we give precedence to the latter.

Even after the application of the C-SMT and context lexicon induction methods, many words remain untranslated. For each such NRL word, we simply check whether it also exists in the RL data, and create the corresponding pair whenever it does. This mainly allows us to supplement our translation lexicon with punctuation signs, but also abbreviations, numbers and proper nouns.

4.4. Evaluation

We evaluate the lexicon induction task on the basis of the dictionaries made available through the Apertium project for the Iberic language pairs (Forcada et al., 2011). Table 2 shows, for each language pair, the number of word pairs covered by the Apertium reference lexicon, the number of word pairs and correctness percentages of the three lexicons obtained by our methods: the C-SMT lexicon (Section 4.1), the context similarity lexicon (Section 4.2) and the merged lexicon (Section 4.3).[12]

Manual evaluation of the SK–CS lexicon yielded comparable correctness scores of above 70%, while the Germanic and the other Slavic language pairs have C-SMT

[12] The merged lexicons evaluated here do not include the identical word pairs.

	Apertium	C-SMT		Contexts		Merged	
	Words	Words	Accuracy	Words	Accuracy	Words	Accuracy
AN–ES	40 469	85 684	75.0%	3 374	88.3%	86 271	75.2%
AST–ES	46 777	69 202	79.7%	7 464	92.8%	70 489	80.1%
CA–ES 500k	105 700	14 378	76.0%	939	93.9%	14 615	76.7%
CA–ES 140M	105 700	678 990	62.5%	20 888	92.5%	681 778	63.8%
GL–ES	76 635	254 594	73.6%	22 853	94.8%	257 413	75.0%
GL–PT	61 388	250 325	58.5%	12 691	87.8%	251 989	59.2%

Table 2: Evaluation of the lexicon induction task on Iberic language pairs. The accuracy is computed on the intersection of the NRL words contained in Apertium and in the respective system.

	Tokens				Types			
	C-SMT	Context	Identical	Suffix	C-SMT	Context	Identical	Suffix
AN–ES	14.3%	49.6%	19.8%	16.3%	12.8%	1.6%	3.3%	82.2%
AST–ES	11.4%	53.9%	18.6%	16.1%	14.1%	3.6%	3.7%	78.6%
CA–ES 500k	18.3%	47.5%	16.0%	18.2%	22.4%	2.6%	6.2%	68.8%
CA–ES 140M	11.1%	57.6%	16.3%	15.0%	3.3%	1.0%	0.8%	94.9%
GL–ES	8.4%	58.6%	18.6%	14.3%	5.9%	2.7%	1.1%	90.3%
GL–PT	14.6%	55.4%	20.6%	9.5%	14.7%	2.2%	3.3%	79.8%
NL–DE	23.0%	27.7%	19.2%	30.1%	17.2%	0.3%	5.5%	77.0%
PFL-DE	21.3%	23.0%	24.2%	31.5%	16.2%	0.4%	8.5%	75.0%
CSB-CS	14.3%	6.97%	29.3%	49.4%	11.3%	0.0%	3.5%	85.1%
DSB-CS	16.0%	11.8%	28.6%	43.6%	12.8%	0.1%	3.5%	83.6%
HSB-CS	18.5%	15.0%	27.7%	38.9%	11.3%	0.3%	2.6%	85.8%
SK-CS	7.6%	46.8%	20.0%	25.7%	4.8%	2.4%	0.9%	91.9%
PL-CS	22.9%	20.7%	22.5%	33.9%	3.6%	0.1%	0.3%	96.1%
CSB-PL	17.0%	14.2%	31.9%	37.0%	15.8%	0.3%	10.3%	73.7%
DSB-PL	18.9%	10.7%	31.9%	38.5%	15.9%	0.1%	6.4%	77.6%
HSB-PL	22.5%	11.8%	30.8%	34.9%	15.2%	0.2%	4.7%	80.0%
SK-PL	26.9%	23.9%	22.7%	26.5%	9.7%	0.2%	1.0%	89.2%
CS-PL	29.3%	23.9%	22.3%	24.5%	7.7%	0.1%	0.7%	91.4%

Table 3: Distribution of the origin of the induced POS tags, by word types and tokens.

correctness scores between 20% and 30% only. These strong discrepancies mostly result from the differences in graphemic compacity and in language proximity among language pairs. They are already visible during the initial cognate pair extraction step and are amplified in the subsequent steps of our pipeline.

5. Creation of the POS-annotated corpus

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. For the time being, we do not deal with potential ambiguities, but rather associate each word unambiguously with the most frequent POS tag of the most frequent translation equivalent. With this simplification, merging the two dictionaries by transitivity is straightforward.

5.2. Adding morphological annotations by suffix analogy

At this point, not all NRL words have been tagged, because some NRL words do not appear in the translation lexicon, or because their RL counterpart are not found in the tag dictionary. In this case, we perform an analogy-based guessing, based on its longest suffix which is also the suffix of a NRL word known to our translation dictionary (in case of ambiguity, we select again the most frequent POS).

5.3. POS annotation and evaluation

Finally, we tag each word in the raw NRL corpus with its unique POS tag. The distribution of the origin of the induced POS tags, both by word type and by token, is given in Table 3.

The tagging accuracy has been evaluated for seven languages on the basis of a manually annotated gold corpus comprising between 30 and 100 sentences per language. The relevant figures, broken down into the different tag induction methods, are shown in Table 4. They show that our approach gives satisfying results, except for Germanic languages which perform worse, the best score being as high as 91.6% (on Slovak, based on Czech annotated data).

5.4. A few words on tagset mismatches

As mentioned above, the deepest limitation of our approach lies in the hypothesis that the inventory of tagsets is the same for both the source RL language and the target NRL language. We were not able to carry out a careful

	C-SMT	Context	Identical	Suffix	Total	# Tags
AN←ES	85.6%	91.2%	91.2%	49.7%	85.4%	42
CA←ES 500k	84.0%	95.9%	96.2%	47.7%	85.9%	42
CA←ES 140M	74.7%	93.6%	99.4%	60.0%	89.1%	42
NL←DE	50.9%	81.7%	78.3%	31.7%	59.0%	55
PFL←DE	72.9%	81.7%	81.9%	35.9%	65.1%	55
HSB←CS	67.1%	93.4%	96.3%	77.2%	83.6%	57
SK←CS	86.8%	94.7%	97.8%	82.0%	91.6%	57
PL←CS	68.8%	85.9%	95.9%	67.0%	77.6%	57

Table 4: Token tagging accuracy. The arrow indicates the RL from which the POS tags were transferred.

qualitative and quantitative analysis of this phenomenon for all language pairs. However, in order to illustrate this phenomenon, let us consider the following example from our manually corrected Polish evaluation corpus:

Nie chcieliśmy rozwiązywać zespołu [...]
'We did not want to dissolve the band [...].'

The two first words, *nie chcieliśmy* 'we didn't want,' are a negative clitic *nie* and a past verb form from the verb *chieć* 'want.' However, such a negative clitic does not exist in Czech, and there is no synthetic form corresponding to *chcieliśmy* either, despite the fact that equivalent morphs are used in the same order for expressing the same thing in both languages. Indeed, Czech and Czech spelling groups together a negative marker *ne* and a past participle form — *nechtěli* — and then adds an auxiliary form *jsme*. As a result, there is no satisfying way to tag neither *nie* (non-existent category in Czech) nor *chcieliśmy* (non-existent verbal sub-category in Czech). In such cases, during the manual development of our evaluation corpora, we tagged such words with non-existent tags: any prediction made by our system therefore counts as an error.

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 unsupervised part-of-speech tagging, for which a POS dictionary of the respective language is generally available. We have applied our methodology to three language sets involving Romance (Iberic), Germanic and Slavic languages.

This work is still to be improved. Our next objective, for example, is to cope with POS ambiguity in a satisfying way, thus paving the way to training POS taggers that perform contextual disambiguation.

Acknowledgements

This work was funded by the Labex EFL (ANR/CGI), Strand 6, operation LR2.2.

7. References

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate production using character-based machine translation. In *Proceedings of IJCNLP 2013*.

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

Anna Feldman, Jirka Hana, and Chris Brew. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of LREC 2006*, pages 549–554.

Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of RANLP 2011)*, 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.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in French and English. In *Proceedings of RANLP 2005)*, 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.

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 ACL 2007 demonstration session*.

Grzegorz Kondrak and Bonnie Dorr. 2004. Identification of confusable drug names: A new approach and evaluation methodology. 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 NAACL 2001*, pages 151–158.

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 ACL 1999*, pages 519–526.

Yves Scherrer and Benoît Sagot. 2013. Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources. In *RANLP 2013 Workshop on Adaptation of language resources and tools for closely related languages and language variants*.

Jörg Tiedemann. 2009. Character-based PSMT for closely related languages. In *Proceedings of EAMT 2009*, pages 12 – 19.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of WMT 2007*, pages 33–39.

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.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT 2001*.