Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora

David Yarowsky
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218 USA
yarowsky@cs.jhu.edu

Grace Ngai
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218 USA
gyn@cs.jhu.edu

Richard Wicentowski
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218 USA
richardw@cs.jhu.edu

ABSTRACT

This paper describes a system and set of algorithms for automatically inducing stand-alone monolingual part-of-speech taggers, base noun-phrase bracketers, named-entity taggers and morphological analyzers for an arbitrary foreign language. Case studies include French, Chinese, Czech and Spanish.

Existing text analysis tools for English are applied to bilingual text corpora and their output projected onto the second language via statistically derived word alignments. Simple direct annotation projection is quite noisy, however, even with optimal alignments. Thus this paper presents noise-robust tagger, bracketer and lemmatizer training procedures capable of accurate system bootstrapping from noisy and incomplete initial projections.

Performance of the induced stand-alone part-of-speech tagger applied to French achieves 96% core part-of-speech (POS) tag accuracy, and the corresponding induced noun-phrase bracketer exceeds 91% F-measure. The induced morphological analyzer achieves over 99% lemmatization accuracy on the complete French verbal system.

This achievement is particularly noteworthy in that it required absolutely no hand-annotated training data in the given language, and virtually no language-specific knowledge or resources beyond raw text. Performance also significantly exceeds that obtained by direct annotation projection.

Keywords

multilingual, text analysis, part-of-speech tagging, noun phrase bracketing, named entity, morphology, lemmatization, parallel corpora

1. TASK OVERVIEW

A fundamental roadblock to developing statistical taggers, bracketers and other analyzers for many of the world's 200+ major languages is the shortage or absence of annotated training data for the large majority of these languages. Ideally, one would like to leverage the large existing investments in annotated data and tools for resource-rich languages (such as English and Japanese) to overcome the annotated resource shortage in other languages.

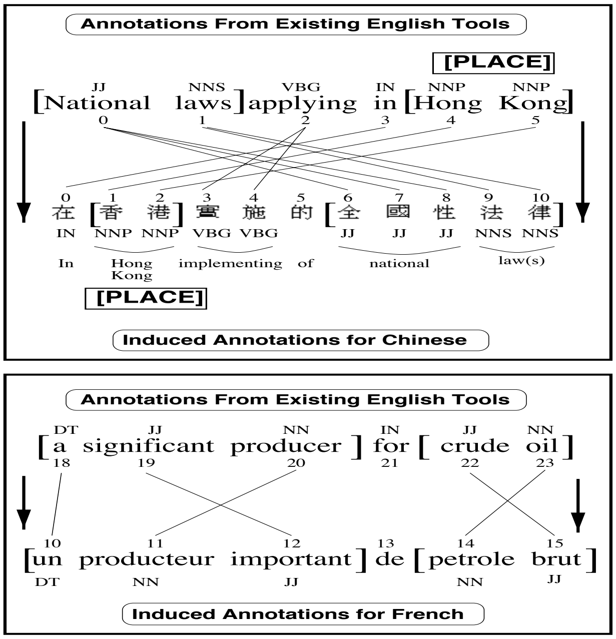

Annotations From Existing English Tools
[PLACE]
JJ
NNS
VBG
IN
NNP
NNP
[National laws] applying in [Hong Kong]
0
1
2
3
4
5
0
1
2
3
4
5
6
7
8
9
10
在 [香 港] 實 施 的 [全 國 性 法 律]
IN
NNP NNP
VBG VBG
JJ
JJ
JJ
NNS NNS
In
Hong Kong
implementing
of
national
law(s)
[PLACE]
Induced Annotations for Chinese
Annotations From Existing English Tools
DT
JJ
NN
IN
JJ
NN
[a significant producer] for [crude oil]
18
19
20
21
22
23
10
11
12
13
14
15
[un producteur important] de [petrole brut]
DT
NN
JJ
NN
JJ
Induced Annotations for French

Figure 1: Projecting part-of-speech tags, named-entity tags and noun-phrase structure from English to Chinese and French.

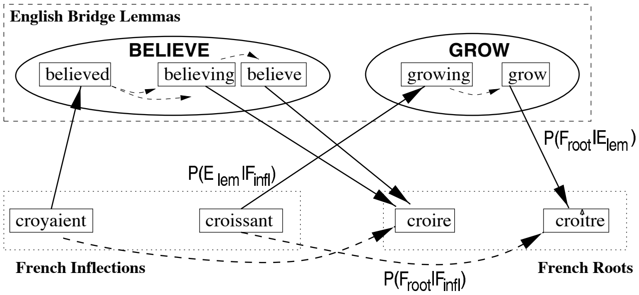


Figure 2: French morphological analysis via English

To show the broad potential of our approach and methods, this paper will investigate four fundamental language analysis tasks: POS tagging, base noun phrase (baseNP) bracketing, named entity tagging, and inflectional morphological analysis, as illustrated in Figures 1 and 2. These bedrock tools are important components of the language analysis pipelines for many applications, and their low cost extension to new languages, as described here, can serve as a broadly useful enabling resource.

2. BACKGROUND

Previous research on the word alignment of parallel corpora has tended to focus on their use in translation model training for MT rather than on monolingual applications. One exception is bilingual parsing. Wu (1995, 1997) investigated the use of concurrent parsing of parallel corpora in a transduction inversion framework, helping to resolve attachment ambiguities in one language by the coupled parsing state in the second language. Jones and Havrilla (1998) utilized similar joint parsing techniques (twisted-pair grammars) for word reordering in target language generation.

However, with these exceptions in the field of parsing, to our knowledge no one has previously used linguistic annotation projection via aligned bilingual corpora to induce traditional standalone monolingual text analyzers in other languages. Thus both our proposed projection and induction methods, and their application to multilingual POS tagging, named-entity classification and morphological analysis induction, appears to be highly novel.

3. DATA RESOURCES

The data sets used in these experiments included the English-French Canadian Hansards, the English-Chinese Hong Kong Hansards, and parallel Czech-English Reader's Digest collection. In addition, multiple versions of the Bible were used, including the French Douay-Rheims Bible, Spanish Reina Valera Bible, and three English Bible Versions (King James, New International and Revised Standard), automatically verse-aligned in multiple pairings. All corpora were automatically word-aligned by the now publicly available EGYPT system (Al-Onaizan et al., 1999), based on IBM's Model 3 statistical MT formalism (Brown et al., 1990). The tagging and bracketing tasks utilized approximately 2 million words in each language, with the sample sizes for morphology induction given in Table 3. All word alignments utilized strictly raw-word-based model variants for English/French/Spanish/Czech and character-based model variants for Chinese, with *no* use of morphological analysis or stemming, POS-tagging, bracketing or dictionary resources.

4. PART-OF-SPEECH TAGGER INDUCTION

Part-of-speech tagging is the first of four applications covered in this paper. The goal of this work is to project POS analysis capabilities from one language to another via word-aligned parallel bilingual corpora. To do so, we use an existing POS tagger (e.g. Brill, 1995) to annotate the English side of the parallel corpus. Then, as illustrated in Figure 1 for Chinese and French, the raw tags are transferred via the word alignments, yielding an extremely noisy initial training set for the 2nd language. The third crucial step is to generalize from these noisy projected annotations in a robust way, yielding a stand-alone POS tagger for the new language that is considerably more accurate than the initial projected tags.

Additional details of this algorithm are given in Yarowsky and Ngai (2001). Due to lack of space, the following sections will serve primarily as an overview of the algorithm and its salient issues.

4.1 Part-of-speech Projection Issues

First, because of considerable cross-language differences in fine-grained tag set inventories, this work focuses on accurately assigning core POS categories (e.g. noun, verb, adverb, adjective, etc.), with additional distinctions in verb tense, noun number and pronoun type as captured in the English tagset inventory. Although impoverished relative to some languages, and incapable of resolving details such as grammatical gender, this Brown-corpus-based tagset granularity is sufficient for many applications. Furthermore, many finer-grained part-of-speech distinctions are resolved primarily by morphology, as handled in Section 7. Finally, if one desires to induce a finer-grained tagging capability for case, for example, one should project from a reference language such as Czech, where case is lexically marked.

Figure 3 illustrates six scenarios encountered when projecting POS tags from English to a language such as French. The first two show straightforward 1-to-1 projections, which are encountered in roughly two-thirds of English words. Phrasal (1-to-N) alignments offer greater challenges, as typically only a subset of the aligned words accept the English tag. To distinguish these cases, we initially assign position-sensitive phrasal parts-of-speech via subscripting (e.g. *Les*/NNS_a *lois*/NNS_b), and subsequently learn a probablistic mapping to core, non-phrasal parts of speech (e.g. $\text{P}(\text{DT}|\text{NNS}_a)$) that is used along with tag sequence and lexical prior models to re-tag these phrasal POS projections.

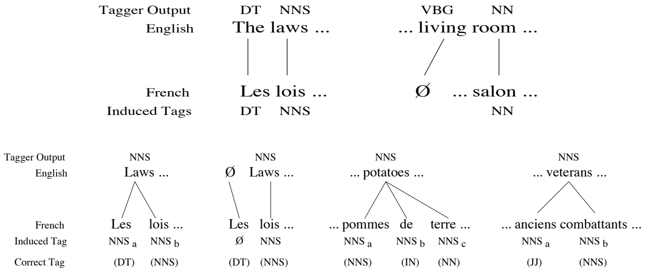


Figure 3: French POS tag projection scenarios

4.2 Noise-robust POS Tagger Training

Even at the relatively low tagset granularity of English, direct projection of core POS tags onto French achieves only 76% accuracy using EGYPT's automatic word alignments (as shown in Table 1). Part of this deficiency is due to word-alignment error; when word alignments were manually corrected, direct projection core-tag accuracy increased to 85%. Also, standard bigram taggers trained on the automatically projected data achieve only modest success at generalization (86% when reapplied to the noisy training data). More highly lexicalized learning algorithms exhibit even greater potential for overmodeling the specific projection errors of this data.

Thus our research has focused on noise-robust techniques for distilling a conservative but effective tagger from this challenging raw projection data. In particular, we modify standard n-gram modeling to separate the training of the tag sequence model $P(T)$ from the lexical prior models $P(W|T)$, and apply different confidence weighting and signal amplification techniques to both.

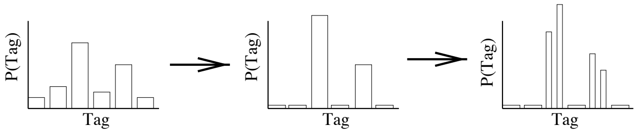
4.2.1 Lexical Prior Estimation

Figure 4 illustrates the process of hierarchically smoothing the lexical prior model $\hat{P}(t|w)$. One motivating empirical observation is that words in French, English and Czech have a strong tendency to exhibit only a single core POS tag (e.g. N or V), and very rarely have more than 2. In English, with relatively high $P(\text{POS}|w)$ ambiguity, only 0.37% of the tokens in the Brown Corpus are not covered by a word type's two most frequent core tags, and in French the percentage of tokens is only 0.03%. Thus we employ an ag-

Model	Evaluate on E-F Aligned French		Evaluate on Unseen Monolingual French	
	Core Tagset	Eng Eqv Tagset	Core Tagset	Eng Eqv Tagset
(a) Direct transfer (on auto-aligned data)	.76	.69	N/A	N/A
(b) Direct transfer (on hand-aligned data)	.85	.78	N/A	N/A
(c) Standard bigram model (on auto-aligned data)	.86	.82	.82	.68
(d) Noise-robust bigram induction (on auto-aligned data)	**.96**	.93	.94	.91
(e) Fully supervised bigram training (on goldstandard)	.97	.96	.98	.97

Table 1: Evaluation of 5 POS tagger induction models on 2 French datasets and 2 tagset granularities

gressive re-estimation in favor of this bias, amplifying the model probability of the majority POS tag, and reducing or zeroing the model probability of 2nd or lower ranked core tags proportional to their relative frequency with respect to the majority tag. This process is then applied recursively, similarly amplifying the probability of the majority subtags within each core tag. Further details, including the handling of *1-to-N* phrasal alignment projections, are given in Yarowsky and Ngai (2001).



Word	Directly Projected Tag					→	Smoothed $\hat{P}(t\|w)$					
	J	N	V	R	I		N	V	NN	NNS	VBN	VBG
achat	0	**62**	48	0	1		.76	.24	.73	.03	.03	.21
cadre	2	**35**	7	1	1		.90	.10	.86	.04	.03	.00
cadres	1	**5**	0	0	0		.94	.00	.04	.90	.00	.00
prévu	1	11	**48**	0	0		.09	.91	.08	.01	.86	.00

Figure 4: Hierarchical smoothing of $\hat{P}(t|w)$ tag probabilities

4.2.2 Tag Sequence Model Estimation

In contrast, the training of the tag sequence model $P(t_i|t_{i-1},...)$ focuses on confidence weighting and filtering of projected training subsequences. The contribution of each candidate training sentence is weighted proportionally with both its EGYPT/GIZA sentence-level alignment score and an agreement measure between the projected tags and the 1st iteration lexical priors, a rough measure of alignment reasonableness. Given the observed bursty distribution of alignment errors in the corpus, this downweighting of low-confidence alignment regions substantially improves sequence model quality with tolerable reduction in training volume.

4.3 Evaluation of POS Tagger Induction

As shown in Table 1, performance is evaluated on two evaluation data sets, including an independent 200K-word hand-tagged French dataset provided by Université de Montréal, which is used to gauge stand-alone tagger performance. Signal amplification and noise reduction techniques yield a 71% error reduction, achieving a core tagset accuracy of 96%, closely approaching the upper-bound 97% performance of an equivalent bigram model trained directly on an 80% subset of the hand-tagged evaluation set (using 5-fold cross-validation). Thus robust training on 500K words of very noisy but automatically-derived tag projections can approach the performance obtained by fully supervised learning on 80K words of hand-tagged training data.

5. NOUN PHRASE BRACKETER INDUCTION

Our empirical studies show that there is a very strong tendency for noun phrases to cohere as a unit when translated between languages, even when undergoing significant internal re-ordering. This strong noun-phrase cohesion even tends to hold for relatively free word order languages such as Czech, where both native speakers and parallel corpus data indicate that nominal modifiers tend to remain in the same contiguous chunk as the nouns they modify. This property allows collective word alignments to serve as a reliable basis for bracket projection as well.

5.1 BaseNP Projection Methodology

The projection process begins by automatically tagging and bracketing the English data, using Brill (1995) and Ramshaw & Marcus (1994), respectively.

As illustrated in Figure 5, each word within an English noun phrase is then subscripted with the number of its NP in the sentence, and this subscript is projected onto the aligned French (or Chinese) words. In the most common case, the corresponding French/Chinese noun phrase is simply the maximal span of the projected subscript.

Figure 6 shows some of the projection challenges encountered. Nearly all such cases of interwoven projected NPs are due to alignment errors, and a strong inductive bias towards NP cohesion was utilized to resolve these incompatible projections.

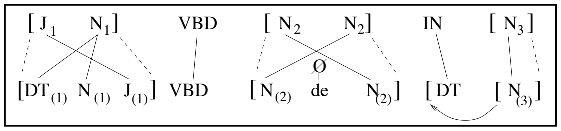


Figure 5: Standard NP projection scenarios.

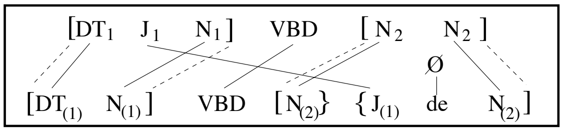


Figure 6: Problematic NP projection scenarios.

5.2 BaseNP Training Algorithm

For stand-alone tool development, the Ramshaw & Marcus IOB bracketing framework and a fast transformation-based learning system (Ngai and Florian, 2001) were applied to the noisy baseNP-projected data described above.

As with POS tagger induction, bracketer induction is improved by focusing training on the highest quality projected data and excluding regions with the strongest indications of word-alignment error. Thus sentences with the lowest 25% of model-3 alignment scores were excluded from training, as were sentences where projected bracketings overlapped and conflicted (also an indicator of alignment errors). Data with lower-confidence POS tagging were not filtered, however, as this filtering reduces robustness when the stand-alone bracketers are applied to noisy tagger output. Additional details are provided in Yarowsky and Ngai (2001).

Current efforts to further improve the quality of the training data include use of iterative EM bootstrapping techniques. Separate projection of bracketings from aligned parallel data with a 3rd language also shows promise for providing independent supervision, which can further help distinguish consensus signal from noise.

5.3 BaseNP Projection Evaluation

Because no bracketed evaluation data were available to us for French or Chinese, a third party fluent in these languages hand-bracketed a small, held-out 40-sentence evaluation set in both languages, using a set of bracketing conventions that they felt were appropriate for the languages. Table 2 shows the performance relative to these evaluation sets, as measured by exact-match bracketing precision (Pr), recall (R) and F-measure (F).

	Exact Match			Acceptable Match		
Method	Pr	R	F	Pr	R	F
Chinese:						
Direct (auto)	.26	.58	.36	.48	.58	.51
Direct (hand)	.47	.61	.53	.86	.86	.86
French:						
Direct (auto)	.43	.48	.45	.60	.58	**.59**
Direct (hand)	.56	.51	.53	.74	.70	.72
FTBL (auto)	.82	.81	.81	**.91**	**.91**	**.91**

Table 2: Performance of BaseNP induction models

It is important to note, however, that many decisions regarding BaseNP bracketing conventions are essentially arbitrary, and agreement rates between additional human judges on these data were measured at 64% and 80% for French and Chinese respectively. Since the translingual projections are essentially unsupervised and have no data on which to mimic arbitrary conventions, it is also reasonable to evaluate the degree to which the induced bracketings are deemed acceptable and consistent with the arbitrary goldstandard (e.g. no crossing brackets). To this end, an additional pool of 3 judges were asked to further adjudicate the differences between the goldstandard and the projection output, annotating such situations as either *acceptable/compatible* or *unacceptable/incompatible.*

Overall, these translingual projection results are quite encouraging. For the Chinese, they are similar to Wu's 78% precision result for translingual-grammar-based NP bracketing, and especially promising given that no word segmentation (only raw characters) were used. For French, the increase from 59% to 91% F-measure for the stand-alone induced bracketer shows that the training algorithm is able to generalize successfully from the noisy raw projection data, distilling a reasonably accurate (and transferable) model of baseNP structure from this high degree of noise.

6. NAMED ENTITY TAGGER INDUCTION

Multilingual named entity tagger induction is based on the extended combination of the part-of-speech and noun-phrase bracketing frameworks. The entity class tags used for this study were FNAME, LNAME, PLACE and OTHER (other entities including organizations). They were derived from an anonymously donated MUC-6 named entity tagger applied to the English side of the French-English Canadian Hansards data.

Initial classification proceeds on a per-word basis, using an aggressively smoothed transitive projection model similar to those described in Section 7. For a given second-language word *FW* and all English words EW_i aligned to it:

$$P(\text{NEclass}_j|\text{FW}) = \sum_i P(\text{NEclass}_j|\text{EW}_i)\, P_a(\text{EW}_i|\text{FW})$$

$$P(\text{PLACE}|\text{Corée}) = P(\text{PLACE}|\text{Korea})\, P_a(\text{Korea}|\text{Corée}) + \ldots$$

The co-training-based algorithm given in Cucerzan and Yarowsky (1999) was then used to train a stand-alone named entity tagger from the projected data. Seed words for this algorithm were those French words that were both POS-tagged as proper nouns and had an above-threshold entity-class confidence from the lexical projection models.

Performance was measured in terms of per-word entity-type classification accuracy on the French Hansard test data, using the 4-class inventory listed above. Classification accuracy of raw tag projections was only 64% (based on automatic word alignment). In contrast, the stand-alone co-training-based tagger trained on the projections achieved 85% classification accuracy, illustrating its effectivess at generalization in the face of projection noise. Notably, most of its observed errors can be traced to entity classification errors from the original English tagger. In fact, when evaluated on the English translation of the French test data set, the English tagger only achieved 86% classification accuracy on this directly comparable data set. It appears that the projection-induced French tagger achieves performance nearly as high as its original training source. Thus further improvements should be expected from higher quality English training sources.

7. MORPHOLOGICAL ANALYSIS INDUCTION

Bilingual corpora can also serve as a very successful bridge for aligning complex inflected word forms in a new language with their root forms, even when their surface similarity is quite different or highly irregular.

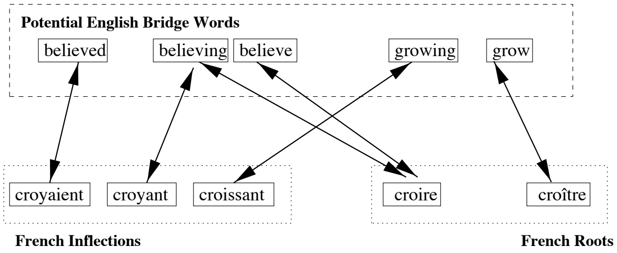


Figure 7: Direct-bridge French inflection/root alignment

As illustrated in Figure 7, the association between a French verbal inflection (*croyant*) and its correct root (*croire*), rather than a similar competitor (*croître*), can be identified by a single-step transitive association via an English bridge word (*believing*). However, in the case of morphology induction, such direct associations are relatively rare given that inflections in a second language tend to associate with similar tenses in English while the singular/infinitive forms tend to associate with analogous singular/infinitive forms, and thus *croyaient* (*believed*) and its root *croire* have no direct English link in our aligned corpus.

However, Figure 2 (first page) illustrates that an existing investment in a lemmatizer for English can help bridge this gap by joining a multi-step transitive association *croyaient*→*believed*→*believe*→ *croire*. Figure 8 illustrates how this transitive linkage via English

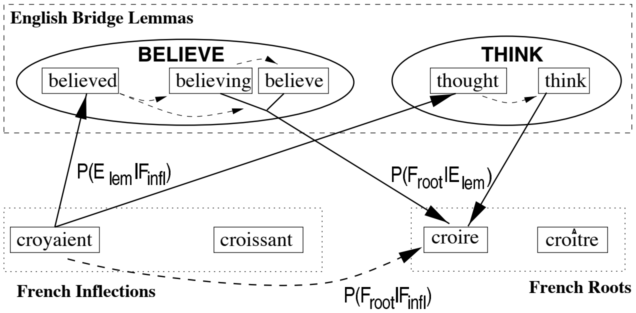


Figure 8: Multi-bridge French inflection/root alignment

lemmatization can be potentially utilized for all other English lemmas (such as THINK) with which *croyaient* and *croire* also associate, offering greater potential coverage and robustness via multiple bridges.

Formally, these multiple transitive linkages can be modeled as shown below, by summing over all English lemmas (E_{lem_i}) with which either a candidate foreign inflection (F_{infl}) or its root (F_{root}) exhibit an alignment in the parallel corpus:

$$P_{mp}(F_{root}|F_{infl}) = \sum_i P_a(F_{root}|E_{lem_i})\, P_a(E_{lem_i}|F_{infl})$$

For example:

$$\begin{aligned} &P_{mp}(\text{croire}|\text{croyaient}) = \\ &\qquad P_a(\text{croire}|\textsc{believe})\, P_a(\textsc{believe}|\text{croyaient}) + \\ &\qquad P_a(\text{croire}|\textsc{think})\, P_a(\textsc{think}|\text{croyaient}) + \ldots \end{aligned}$$

This projection/bridge-based similarity measure $P_{mp}(F_{root}|F_{infl})$ can be quite effective on its own, as shown in the *MProj only* entries in Table 3 (for multiple parallel corpora in 3 different languages), especially when restricted to the highest-confidence subset of the vocabulary (5.2% to 77.9% in these data) for which the association exceeds simple fixed probability and frequency thresholds. When estimated using a 1.2 million word subset of the French Hansards, for example, the MProj measure alone achives 98.5% precision on 32.7% of the inflected French verbs in the corpus (constituting 97.6% of the tokens in the corpus). Unlike traditional string-transduction-based morphology induction methods where irregular verbs pose the greatest challenges, these typically high-frequency words are often the *best* modelled data in the vocabulary making these multilingual projection techniques a natural complement to existing models.

7.1 Trie-based Morphology Models

The high precision on the MProj-covered subset also make these partial pairings effective training data for robust supervised algorithms that can generalize the string transformation behavior to the remaining uncovered vocabulary. While any supervised morphological analysis technique is possible here, we employ a trie-based modeling technique where the probability of a given stem-change (from the inventory observed in the MProj-paired training data) is modeled hierarchically using variable suffix context, as described in Yarowsky and Wicentowski (2000):

$$\begin{aligned} &P(\text{root}|\text{inflection}) = P(\delta\beta|\delta\alpha) = P(\alpha \rightarrow \beta|\delta\alpha) = \\ &\qquad \sum_i \lambda_i P(\alpha \rightarrow \beta|h_i) \quad \text{for } h_i = \text{suffix}(i, \delta\alpha) \end{aligned}$$

For example:

$$\begin{aligned} &P(\text{commencer}|\text{commença}) = P(\text{ça} \rightarrow \text{cer}|\text{commença}) = \\ &\quad \lambda_0 P(\text{ça} \rightarrow \text{cer}) + \lambda_1 P(\text{ça} \rightarrow \text{cer}|\text{a}) + \lambda_2 P(\text{ça} \rightarrow \text{cer}|\text{ça}) + \\ &\quad + \lambda_3 P(\text{ça} \rightarrow \text{cer}|\text{nça}) + \lambda_4 P(\text{ça} \rightarrow \text{cer}|\text{ença}) + \ldots \end{aligned}$$

$$\begin{aligned} &P(\text{ployer}|\text{ploie}) = P(\text{ie} \rightarrow \text{yer}|\text{ploie}) = \\ &\quad \lambda_0 P(\text{ie} \rightarrow \text{yer}) + \lambda_1 P(\text{ie} \rightarrow \text{yer}|\text{e}) + \lambda_2 P(\text{ie} \rightarrow \text{yer}|\text{ie}) + \\ &\quad + \lambda_3 P(\text{ie} \rightarrow \text{yer}|\text{oie}) + \lambda_4 P(\text{ie} \rightarrow \text{yer}|\text{loie}) + \ldots \end{aligned}$$

An important property of the trie-based models is their effectiveness at clustering words that exhibit similar morphological behavior, both reducing model size and facilitating generalization to previously unseen examples. This property is illustrated in Figure 9, showing a sample (inflection → root) trie branch for French verbal inflections, with suffix histories h='oie', h='noie', h='roie', etc. At each history node, the hierarchically smoothed probabilities of several $\alpha \rightarrow \beta$ (inflection→root) changes are given. Note that the relative probabilities of the competing analyses *ie→ir* and *ie→yer* differ substantially for diffent suffix histories, and that there are subexceptions that tend to cluster by affix history. This allows for the successful analysis of 8 of the 9 italicized test words that had not been seen in the bilingual projection data or where the MProj model yielded no root candidate above threshold.

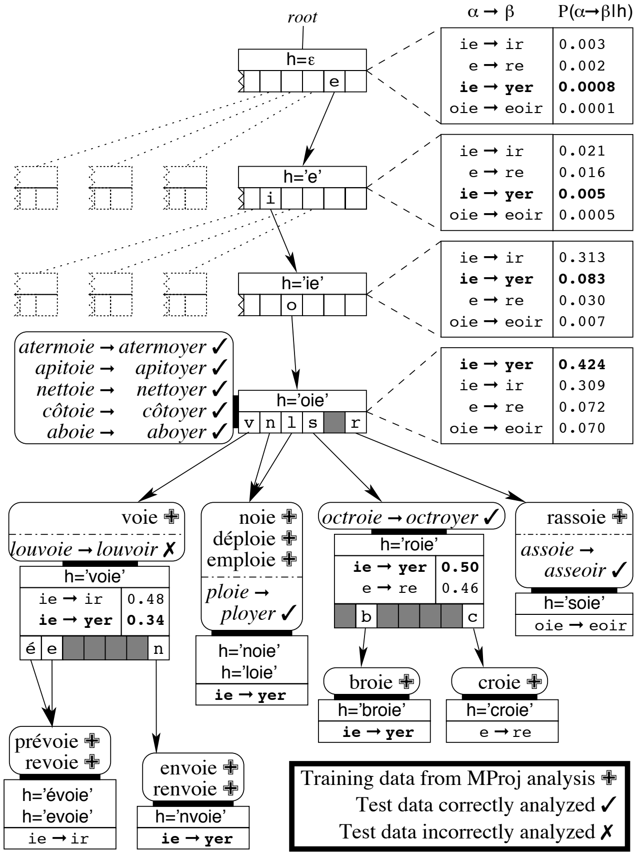


Figure 9: Example of a French MTrie branch, showing inflection → root probabilities ($P(\alpha \rightarrow \beta|h_i)$) for variable length suffix histories (h_i). MTrie analyses on test data are given in italics.

Table 3 illustrates the performance of a variety of morphology induction models. When using the projection-based MProj and trie-based MTrie models together (with the latter extending coverage to words that may not even appear in the parallel corpus), full verb lemmatization precision on the 1.2M word Hansard subset exceeds 99.5% (by type) and 99.9% (by token) with 95.8% coverage by type and 99.8% coverage by token. A backoff model based on Levenshtein-distance and distributional context similarity handles the relatively small percentage of cases where MProj and MTrie together are not sufficiently confident, bringing the system coverage to 100% coverage with a small drop in precision to 97.9% (by type) and 99.8% (by token) on the unrestricted space of inflected verbs observed in the full French Hansards. As shown in Section 7.3, performance is strongly correlated with size of the initial aligned bilingual corpus, with a larger Hansard subset of 12M words yielding 99.4% precision (by type) and 99.9% precision (by token). Performance on Czech is discussed in Section 7.3.

	Precision		Coverage	
Model	Typ	Tok	Typ	Tok
FRENCH Verbal Morphology Induction				
French Hansards (12M words):				
MProj only	.992	.999	.779	.994
MProj+MTrie	.998	.999	.988	.999
MProj+MTrie+BKM	**.994**	**.999**	1.00	1.00
French Hansards (1.2M words):				
MProj only	.985	.998	.327	.976
MProj+MTrie	.995	.999	.958	.998
MProj+MTrie+BKM	**.979**	**.998**	1.00	1.00
French Hansards (120K words):				
MProj only	.962	.931	.095	.901
MProj+MTrie	.984	.993	.916	.994
MProj+MTrie+BKM	**.932**	**.989**	1.00	1.00
French Bible (300K words) via 1 English Bible:				
MProj only	1.00	1.00	.052	.747
MProj+MTrie	.991	.998	.918	.992
MProj+MTrie+BKM	**.954**	**.994**	1.00	1.00
French Bible (300K words) via 3 English Bibles:				
MProj only	.928	.975	.100	.820
MProj+MTrie	.981	.991	.931	.990
MProj+MTrie+BKM	**.964**	**.991**	1.00	1.00
CZECH Verbal Morphology Induction				
Czech Reader's Digest (500K words):				
MProj only	.915	.993	.152	.805
MProj+MTrie	.916	.917	.893	.975
MProj+MTrie+BKM	**.878**	**.913**	1.00	1.00
SPANISH Verbal Morphology Induction				
Spanish Bible (300K words) via 1 English Bible:				
MProj only	.973	.935	.264	.351
MProj+MTrie	.988	.998	.971	.967
MProj+MTrie+BKM	**.966**	**.985**	1.00	1.00
Spanish Bible (300K words) via French Bible:				
MProj only	.980	.935	.722	.765
MProj+MTrie	.983	.974	.986	.993
MProj+MTrie+BKM	**.974**	**.968**	1.00	1.00
Spanish Bible (300K words) via 3 English Bibles:				
MProj only	.964	.948	.468	.551
MProj+MTrie	.990	.998	.978	.987
MProj+MTrie	**.976**	**.987**	1.00	1.00

Table 3: Performance of full verbal morphological analysis, including precision/coverage by type/token

7.2 Morphology Induction via Aligned Bibles

Performance using even small parallel corpora (e.g. a 120K subset of the French Hansards) still yields a respectable 93.2% (type) and 98.9% (token) precision on the verb-lemmatization test set for the full Hansards. Given that the Bible is actually larger (approximately 300K words, depending on version and language) and available on-line or via OCR for virtually all languages (Resnik et al., 2000), we also conducted several experiments on Bible-based morphology induction, further detailed in Table 3.

7.2.1 Boosting Performance via Multiple Parallel Translations

Even though at most one translation of the Bible is typically available in a given foreign language, numerous English Bible versions are freely available and a performance increase can be achieved by simultaneously utilizing alignments to each English version. As illustrated in Figure 10, different aligned Bible pairs may exhibit (or be missing) different full or partial bridge links for a given word (due both to different lexical usage and poor textual parallelism in some text-regions or version pairs). However, $P_a(F_{root}|E_{lem_i})$ and $P_a(E_{lem_i}|F_{\mathit{infl}})$ need not be estimated from the same Bible pair. Even if one has only one Bible in a given source language, each alignment with a distinct English version gives new bridging opportunities with no additional resources needed on the source language side. The baseline approach (evaluated here) is simply to concatenate the different aligned versions together. While word-pair instances translated the same way in each version will be repeated, this rather reasonably reflects the increased confidence in this particular alignment. An alternate model would weight version pairs differently based on the otherwise-measured translation faithfulness and alignment quality between the version pairs. Doing so would help decrease noise. Increasing from 1 to 3 English versions reduces the type error rate (at full coverage) by 22% on French and 28% on Spanish with no increase in the source language resources.

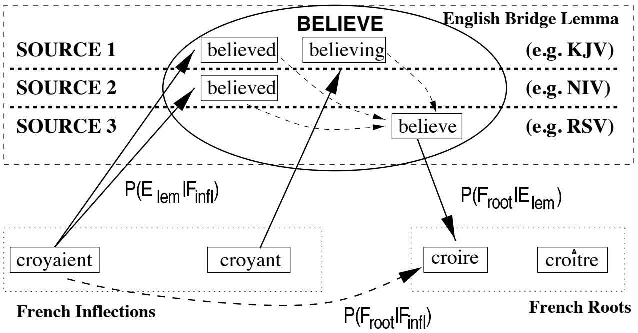


Figure 10: Use of multiple parallel Bible translations

7.2.2 Boosting Performance via Multiple Bridge Languages

Once lemmatization capabilities have been successfully projected to a new language (such as French), this language can then serve as an additional bridging source for morphology induction in a third language (such as Spanish), as illustrated in Figure 11. This can be particularly effective if the two languages are very similar (as in Spanish-French) or if their available Bible versions are a close translation of a common source (e.g. the Latin Vulgate Bible). As shown in Table 3, using the previously analyzed French Bible as a bridge for Spanish achieves performance (97.4% precision) comparable to the use of 3 parallel English Bible versions.

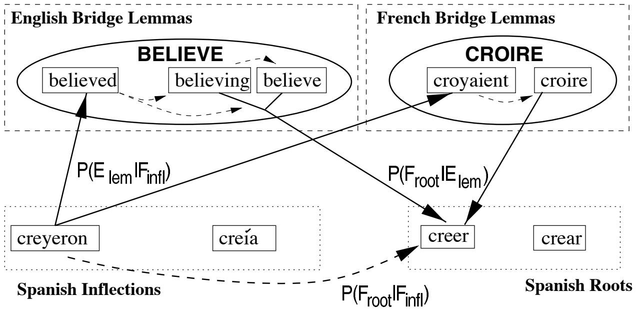


Figure 11: Use of bridges in multiple languages.

7.3 Morphology Induction: Observations

This section includes additional detail regarding the morphology induction experiments, supplementing the previous details and analyses given in Section 7 and Table 3.

- Performance induction using the French Bible as the bridge source is evaluated using the full test verb set extracted from the French Hansards. The strong performance when trained only using the Bible illustrates that even a small single text in a very different genre can provide effective transfer to modern (conversational) French. While the observed genre and topic-sensitive vocabulary differs substantially between the Bible and Hansards, the observed inventories of stem changes and suffixation actually have large overlap, as do the set of observed high-frequency irregular verbs. Thus the inventory of morphological phenomena seem to translate better across genre than do lexical choice and collocation models.
- Over 60% of errors are due to gaps in the candidate rootlists. Currently the candidate rootlists are derived automatically by applying the projected POS models and selecting any word with the probability of being an uninflected verb greater than a generous threshold and also ending in a canonical verb suffix. False positives are easily tolerated (less than 5% of errors are due to spurious non-root competitors), but with missing roots the algorithms are forced either to propose previously unseen roots or align to the closest previously observed root candidate. Thus while *no* non-English dictionary was used in the computation of these results, it would substantially improve performance to have a dictionary-based inventory of potential roots, increasing coverage and decreasing noise from competing non-roots and spelling errors.
- Performance in all languages has been significnatly hindered by low-accuracy parallel-corpus word-alignments using the original Model-3 GIZA tools. Use of Och and Ney's recently released and enhanced GIZA++ word-alignment models (Och and Ney, 2000) should improve performance for all of the applications studied in this paper, as would iterative re-alignments using richer alignment features (including lemma and part-of-speech) derived from this research.
- The current somewhat lower performance on Czech is due to several factors. They include (a) very low accuracy initial word-alignments due to often non-parallel translations of the Reader's Digest sample and the failure of the initial word-alignment models to handle the highly inflected Czech morphology. (b) the small size of the Czech parallel corpus (less than twice the length of the Bible). (c) the common occurrence in Czech of two very similar perfective and non-perfective root variants (e.g. *odolávat* and *odolat*, both of which mean *to resist*). A simple monolingual dictionary-derived list of canonical roots would resolve ambiguity regarding which is the appropriate target.
- Many of the errors are due to all (or most) inflections of a single verb mapping to the same incorrect root. But for many applications where the function of lemmatization is to cluster equivalent words (e.g. stemming for information retrieval), the choice of label for the lemma is less important than correctly linking the members of the lemma.

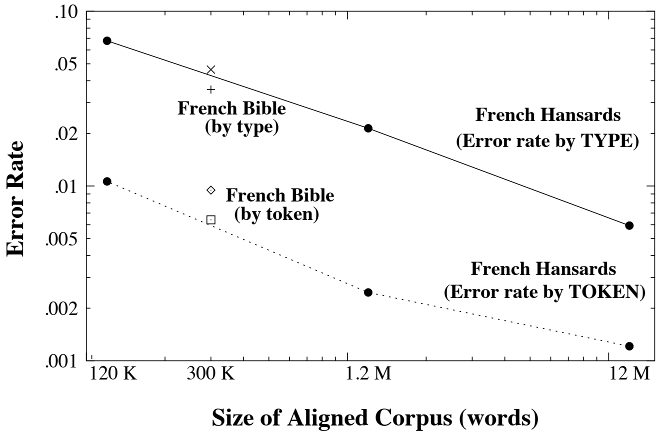


Figure 12: Learning Curves for French Morphology

- The learning curves in Figure 12 show the strong correlation between performance and size of the aligned corpus. Given that large quantities of parallel text currently exist in translation bureau archives and OCR-able books, not to mention the increasing online availability of bitext on the web, the natural growth of available bitext quantities should continue to support performance improvement.
- The system analysis examples shown in Table 4 are representative of model performance and are selected to illustrate the range of encountered phenomena. All system evaluation is based on the task of selecting the correct root for a given inflection (which has a long lexicography-based consensus regarding the "truth"). In contrast, the descriptive analysis of any such pairing is very theory dependent without standard consensus. The "TopBridge" column shows the strongest English bridge lemma utilized in mapping (typically one of many potential bridge lemmas).

These results are quite impressive in that they are based on essentially no language-specific knowledge of French, Spanish or Czech. In addition, the multilingual bridge algorithm is surface-form independent, and can just as readily handle obscure infixational or reduplicative morphological processes.

8. CONCLUSION

This paper has presented a detailed survey of original algorithms for cross-language annotation projection and noise-robust tagger induction, evaluated on four diverse applications. It shows how previous major investments in English annotated corpora and tool development can be effectively leveraged across languages, achieving accurate stand-alone tool development in other languages without comparable human annotation efforts. Collectively this work is the most comprehensive existing exploration of a very promising new paradigm for cross-language resource projection.

Acknowledgements

This research has been partially supported by NSF grant IIS-9985033 and ONR/MURI contract N00014-01-1-0685. The authors thank Silviu Cucerzan, Radu Florian, Jan Hajic, Gideon Mann and Charles Schafer for their valuable contributions and feedback.

9. REFERENCES

[1] Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, FJ Och, D. Purdy, N. Smith and D. Yarowsky. 1999. *Statistical Machine Translation* (tech report). Johns Hopkins University.

[2] E. Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 24(1): 543–565.

[3] P. Brown, J. Cocke, S. DellaPietra, V. DellaPietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Rossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):29–85.

[4] S. Cucerzan and D. Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence." In *Proceedings, 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pp. 90-99.

[5] P. Fung and K. Church. 1994. K-vec: a new approach for aligning parallel texts. In *Proceedings of COLING-94*, pp. 1096–1102.

[6] P. Fung and K. McKeown. 1994. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic warping. In *Proceedings of AMTA-94*, pp. 81–88.

[7] D. Jones, and R. Havrilla. 1998 Twisted pair grammar: Support for rapid development of machine translation for low density languages In *Procs. of AMTA'98*, pp. 318–332.

[8] D. Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

[9] G. Ngai and R. Florian. 2001. Transformation-based learning in the fast lane. In *Proceedings of NAACL-2001*, pp. 40-47.

[10] F.J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of ACL-2000*, pp. 440-447.

[11] L. Ramshaw and M. Marcus, 1999. Text chunking using transformation-based learning. In Armstrong et al. (Eds.), *Natural Language Processing Using Very Large Corpora*. Kluwer, pp. 157-176.

[12] P. Resnik, M. Olsen, and M. Diab. 2000. The Bible as a parallel corpus: annotating the 'Book of 2000 Tongues' *Computers and the Humanities*, 33(1-2):129-153.

[13] D. Wu. 1995. An algorithm for simultaneously bracketing parallel texts. In *Proc. of ACL-95*, pp. 244–251.

[14] D. Wu. 1997. Statistical inversion transduction grammars an bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377-404.

[15] D. Yarowsky and G. Ngai. 2001. Inducing multilingual POS taggers and NP Bracketers via robust projection across aligned corpora. In *Proceedings of NAACL-2001*, pp. 377-404.

[16] D. Yarowsky and R. Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of ACL-2000*, pp. 207-216.

Induced Morphological Analyses for CZECH

Inflection	Root Out	Analysis	TopBridge
bral	brát	al→át	marry
brala	brát	ala→át	accept
brali	brát	ali→át	marry
byl	být	yl→ýt	be
byli	být	yli→ýt	be
bylo	být	ylo→ýt	be
chovala	chovat	la→t	behave
chová	chovat	á→at	behave
chováme	chovat	áme→at	behave
chodila	chodit	la→t	walk
chodí	chodit	í→it	walk
choďte	chodit	ďte→dit	swim
chránila	chránit	la→t	protect
chrání	chránit	í→it	protect
couval	couvat	l→t	back
chce	chtít	ce→tít	want
chcete	chtít	cete→tít	want
chceš	chtít	ceš→tít	want
chci	chtít	ci→tít	want
chtějí	chtít	ějí→ít	want
chtěli	chtít	ěli→ít	want
chtělo	chtít	ělo→ít	want

Induced Morphological Analyses for SPANISH

Inflection	Root Out	Analysis	TopBridge
aborreció	aborrecer	ió→er	hate
aborrecía	aborrecer	ía→er	hate
aborrezco	aborrecer	zco→cer	hate
abrace	abrazar	ce→zar	embrace
abrazado	abrazar	ado→ar	embrace
adquiere	adquirir	ere→rir	get
andamos	andar	amos→ar	walk
andando	andar	ando→ar	walk
andarán	andar	arán→ar	wander
andarás	andar	arás→ar	wander
andemos	andar	emos→ar	walk
anden	andar	en→ar	walk
anduvo	andar	uvo→ar	walk
buscáis	buscar	áis→ar	seek
buscó	buscar	ó→ar	seek
busque	buscar	que→car	seek
busqué	buscar	qué→car	seek

Induced Morphological Analyses for FRENCH

Inflection	Root Out	Analysis	TopBridge
abrège	abréger	ège→éger	shorten
abrègent	abréger	ègent→éger	shorten
abrégerai	abréger	erai→er	curtail
achète	acheter	ète→eter	buy
achètent	acheter	ètent→eter	buy
achètera	acheter	ètera→eter	buy
advenait	advenir	ait→ir	happen
advenu	advenir	u→ir	happen
adviendrait	advenir	iendrait→enir	happen
advient	advenir	ient→enir	happen
aliène	aliéner	ène→éner	alienate
aliènent	aliéner	ènent→éner	alienate
conçu	concevoir	çu→cevoir	conceive
crois	croire	s→re	believe
croyaient	croire	yaient→ire	believe

Table 4: Sample of induced morphological analyses