Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora

David Yarowsky[†] and **Grace Ngai**[†,‡]
{yarowsky,gyn}@cs.jhu.edu

[†] Johns Hopkins University
Baltimore, MD 21218, USA

[‡] Weniwen Technologies
Hong Kong

Abstract

This paper investigates the potential for projecting linguistic annotations including part-of-speech tags and base noun phrase bracketings from one language to another via automatically word-aligned parallel corpora. First, experiments assess the accuracy of unmodified direct transfer of tags and brackets from the source language English to the target languages French and Chinese, both for noisy machine-aligned sentences and for clean hand-aligned sentences. Performance is then substantially boosted over both of these baselines by using training techniques optimized for very noisy data, yielding 94-96% core French part-of-speech tag accuracy and 90% French bracketing F-measure for stand-alone monolingual tools trained without the need for any human-annotated data in the given language.

1 Introduction and 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. Furthermore, hand-annotation of even reasonably well-understood features such as part-of-speech tags and basic phrase structure has proved to be labor intensive and costly. For example, many person years and over $1 million have been invested in the English Penn Treebank alone, and the small minority of languages with currently developed treebanks and tagged corpora indicate that government or private investment may be difficult to raise for annotation projects in most languages. In contrast, the explosive growth of multilingual government and commercial websites and news streams, and the potentially large future market in archived human translations of documents and electronic books suggest that *unannotated* parallel text data is likely to become broadly available.

Ideally, one would like to leverage the major investments in annotated data and tools for resource-rich languages (such as English, French and Japanese) to overcome the annotated resource shortage in other languages. This paper investigates a very promising approach to doing so: use automatically word-aligned raw bilingual corpora to project annotations from a resource-rich language to an arbitrary second language, and develop robust techniques to train on and improve this potentially very noisy induced projection onto of the second language.

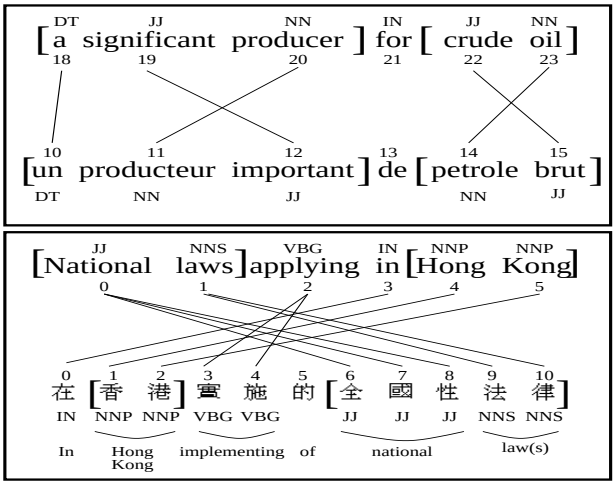

Figure 1: Projecting POS tag and noun-phrase structure across languages (system output).

To illustrate the broad potential of this work, the paper will investigate two quite different tasks: lexical part-of-speech (POS) tagging and structural base noun phrase (BaseNP) bracketing. It will also study two very different languages: French and Chinese. While certainly not resource-poor, these languages were chosen specifically because they had existing annotated corpora on which to evaluate and compare performance. Their diversity also demonstrates the generalizability of these techniques.

Graphically, the projection of POS tags and NP structure is shown in Figure 1. Both examples are actual outputs of the algorithms described below. All steps in this process are completely automatic, including the POS tagging and bracketing on the English side, the word-alignment between languages and the induced annotations on the Chinese and

French sides.

There are two central limitations to this paradigm, however. The first is the often very poor accuracy of word alignments, due both to the current limitations of word-alignment algorithms, and also to the often weak or incomplete inherent match between the two sides of a bilingual corpus. The paper will address and handle this problem through robust, noise-tolerant learning algorithms capable of being trained effectively on incomplete and highly inaccurate alignments. The second limitation is the potential mismatch in the annotation needs of two languages; not all distinctions that may be desirable for one language (such as grammatical gender in French) are compatible or even present in a parallel language such as English. The paper will discuss solutions to these language-level mismatches, and will illustrate that at the level of noun-phrase structure and core part-of-speech tags, essential annotations can be projected with remarkable effectiveness and coverage in many cases.

Finally, the paper will empirically evaluate two major questions for each of the tasks:

- The accuracy of the direct projections of BaseNP structure and POS tags across languages when (a) word alignments are derived fully automatically (with heavy noise), and (b) word alignments are hand-corrected and as optimal as possible. The latter offers an upper bound for direct transfer accuracy.
- The algorithms' ability to generalize from the noisy training data and tag a held-out monolingual corpus, (c) when standard algorithms are applied directly to the noisy data without modification, and (d) when the robust algorithms described below are employed. The high accuracy of the latter, significantly outperforming direct transfer on cleanly aligned data, indicates the importance of the induction algorithm beyond simple projection, even under ideal circumstances.

2 Background

The approach and general algorithms investigated in this paper were initiated in conjunction with the EGYPT project of the 1999 Johns Hopkins summer machine translation workshop (Al-Onaizan et al., 1999). Previously, tools for automatic word-alignment of bilingual corpora were not widely available outside IBM, the research group pioneering statistical machine translation with the Candide system (Brown et al, 1990). The researchers who developed independent word-alignment tools (e.g. Dagan et al, 1993; Fung and Church, 1994; Wu, 1994; Melamed, 1999; Och and Ney, 2000) tended to focus on translation model applications for their word-alignments rather than the induction of stand-alone *monolingual* analyzers via cross-language projection. For example, Kupiec (1993) began with existing Xerox monolingual bracketers to improve translation alignments, rather than the converse.

The primary exception has been in the area of parallel bilingual parsing. Wu (1995, 1997) proposed a framework for inversion transduction grammars, where parallel corpora in languages such as English and Chinese are parsed concurrently, with cross-language order differences captured via mobile-like CFG production reordering. Structural relationships in one language help constrain structural relationships in the second language. Evaluation on noun-phrase bracketing showed 78% precision for Chinese, and 80% precision for English. Thus, while remarkably effective for learning without human-annotated training data, the algorithm does assume the existence of a parallel second-language mirror for all sentences to be parsed. Also, Wu observed significant performance degradation when either the word alignment or translation faithfulness in these pairs are weak. This further motivates the noise-robust training and stand-alone application of our current work.

In a related framework, Jones and Havrilla (1998) investigated the use of twisted-pair grammars for syntactic transfer. Given an existing Hindi/Urdu sentence parse, English output was generated by rotating subtrees using the constraints and preferences of the transduction grammar. The ability to generate candidate target-language orderings in this manner offers great potential to productively constrain search in a statistical MT system. Yet the assumption of existing syntactic analyses for each source language further motivates the need to induce such analyses.

3 Data Resources

The data used in our experiments are the English-French Canadian Hansards and English-Chinese Hong Kong Hansards, parallel records of parliamentary proceedings and publications. Both corpora were word-aligned by the now publicly available EGYPT system (Al-Onaizan et al., 1999) and based on IBM's Model 3 statistical MT formalism (Brown et al., 1990). The data sets used for our projection studies both contained approximately 2 million words in each language. Their alignment was based on strictly word-based model variants for English and character-based model variants for Chinese, with *no* use of morphological analysis or stemming, POS-tagging, bracketing, outside dictionaries or any other external data source or annotation tool.[1] Thus the experiments were carefully designed

[1] The two exceptions are end-of-sentence detection and tokenization. For the French Hansards, before alignment only 4 simple transformations were performed: au→a le, aux→a

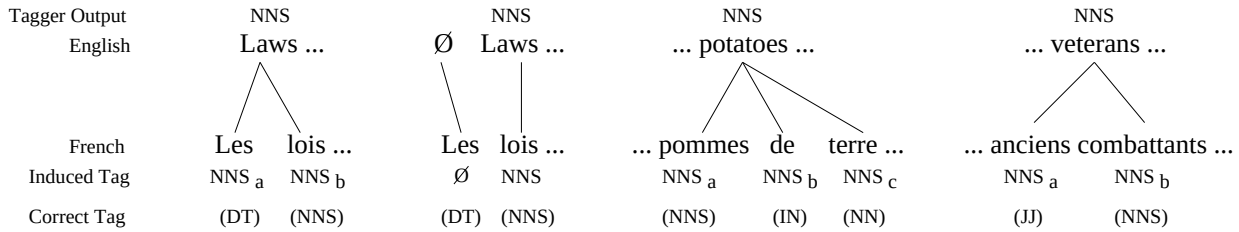

Figure 2: Problematic English-French tag projection scenarios.

not to depend on any analysis or annotation that they intended to induce.

While the French and Chinese Hansards were used in this paper because they are standard reference sets for evaluation purposes, the algorithms studied here do not depend on these particular languages or corpora. A multitude of resources are (or are becoming) available, including MULTEXT-East corpus of Orwell's 1984, the Bible (available in most languages), loosely parallel online news bitexts, and the surprisingly underutilized bitext archives of commercial translation services. Also, while English has been used here as the projection source language, other languages with existing annotation tools and appropriate parallel corpora could readily substitute in this role.

4 Part-of-Speech Tagger Induction

Part-of-speech tagging is the first of the applications covered in this paper. The goal of this work is to use an existing POS tagger for English (e.g. Brill, 1995) to annotate the English side of a parallel corpus, then project the tag annotations to the second language, and then generalize from this noisy projection in a robust way.

As previously noted, any two languages will exhibit tagset mismatches and differences in their morphologically realized POS granularity. These issues are discussed further in Section 4.4, and the intervening sections will assume that the goal of this study is to reliably project onto a second language (e.g. French) text the level of POS granularity realized in the English Penn Treebank tagset, such as NN and NNS for singular and plural nouns (but not finer distinctions of grammatical gender) and basic Treebank verb tenses VB/VBN/VBG/VBD (but not the more subtle simple-past/imperfect tense distinctions, mood or person/number differences that are morphologically realized in French).

4.1 Initial Direct POS Projection

Figures 2 and 3 illustrate several scenarios in the projection of English POS tags across IBM Model 3

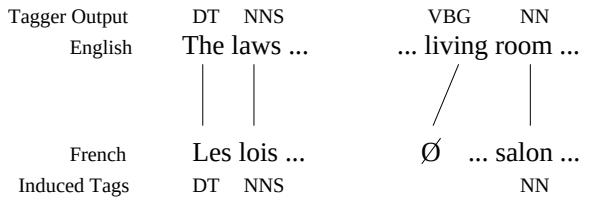

Figure 3: Simplest E-F tag projection case.

word-alignments to French (assuming English as the noisy channel "source").

Figure 3 shows the ideal situation where alignments are 1-to-1, and thus tend to be relatively clean POS projections. However, an important artifact of this alignment model and direction is that while each French word token corresponds to exactly 1 English token, a given English token may correspond to many French tokens. Thus a determiner-free use of the English *Laws* and *potatoes* may correspond (as in Figure 2) to the phrases *Les lois* and *pommes de terre* respectively. Which (or all) of these French words should inherit the English plural noun (NNS) tag?

One option is to project the English tag only to the French word token where the alignment probability is highest. This is problematic in that it can be difficult to extract from Model 3 tools, and because doing so will leave numerous French word tokens without any projected tag. An alternate approach, pursued here, is to subscript the projection of a tag onto multiple French words in a compound with the relative position in the compound (a, b, c, etc.). Thus NNS_a, corresponding to the first 1-to-n alignment position in a French compound, will tend to have a high probability of corresponding to a French determiner, while NNS_b, in second position, will tend to have a low probability of corresponding to a determiner. Distinguishing 1-to-1 projections and each position in 1-to-n projections allows these different cases to be modeled separately.

4.2 Robust Learning from Noisy Tag Projections

Unfortunately, as shown in Section 4.3, English-to-French tag projections exhibit considerable noise,

les, du→de le, and des→de les, using basic context heuristics utilized in the off-the-shelf EGYPT distribution.

even when the high-error automatic alignments have been manually corrected, yielding 69% and 78% direct projection accuracy respectively (at English tagset granularity). Traditional supervised learning algorithms tend to perform poorly at this level of noise, and a standard bigram tagger trained on the automatically aligned (uncorrected) data achieves only 82% when evaluated on a held-out test set. 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. To do so, we (a) downweight or exclude training data segments identified as poorly aligned or likely noise (b) use a conservative bigram learning algorithm, and (c) train the lexical prior and tag-sequence models separately using aggressive generalization techniques.

4.2.1 Lexical Prior Estimation

In a standard bigram tagging model, one selects a tag sequence T for a word sequence W by:

$$\underset{T}{\text{argmax}} \quad P(T|W) = P(W|T)P(T)$$

where

$$P(T) = P(t_1....t_n) \approx P(t_1)P(t_2|t_1)..P(t_n|t_{n-1})$$

and

$$P(W|T) = P(w_1...w_n|t_1...t_n) \approx \prod_{i=1}^{n} P(w_i|t_i)$$

using standard independence assumptions. Section 4.2.2 will discuss the estimation of $P(t_i|t_{i-1})$. The following section describes the estimation of $P(t_i|w_i)$, which using Bayes rule and direct (relatively noise-free) measurement of $P(w_i)$ from the French data, can be used to calculate $P(w_i|t_i)$ as:

$$P(w_i|t_i) = \frac{P(t_i|w_i)P(w_i)}{\sum_j P(t_i|w_j)P(w_j)}$$

Inspection of the raw projected tag data shows the need for an improved estimation of $P(t|w)$. Temporarily excluding the case of compound alignments (e.g. NNS_a), Table 1 shows the observed frequency distributions of English tags projected onto four French words from 1-to-1 alignments, for the core N/V/J/R/I POS tags. Note that the total probability mass assigned to potentially correct tags (in bold) is relatively low, with fairly broad misassignment to incorrect tags for the given word.

At the core tag level in particular, we observe empirically that words in French have a strong tendency to have only 1 possible core POS tag, and very rarely have more than 2. Even 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 drops to 0.03%. Thus we employ an aggressive re-estimation in favor of this bias, where for $t_{(i)}$ = the i^{th} most frequent tag for w:

$$\hat{P}(t_{(2)}|w) = \lambda_1 P(t_{(2)}|w) \qquad \text{where } \lambda_1 < 1.0$$
$$\hat{P}(t_{(1)}|w) = 1 - \hat{P}(t_{(2)}|w)$$
$$\hat{P}(t_{(c)}|w) = 0 \qquad \text{for all } c > 2$$

giving the large majority of the new probability mass to the single highest frequency core tag.

	Directly Projected Tag					Tag
Word	J	N	V	R	I	Error
achat	0	**62**	48	0	1	0.44
cadre	2	**35**	7	1	1	0.27
cadres	1	**5**	0	0	0	0.17
prévu	1	11	**48**	0	0	0.20

Table 1: Raw projected tag distributions.

	Smoothed $\hat{P}(t\|w)$					
Word	N	V	NN	NNS	VBN	VBG
achat	.76	.24	.73	.03	.03	.21
cadre	.90	.10	.86	.04	.03	.00
cadres	.94	.00	.04	.90	.00	.00
prévu	.09	.91	.08	.01	.86	.00

Table 2: Smoothed $\hat{P}(t|w)$ tag probabilities

Applying this model recursively, the finer grained subtag probabilities (e.g. NN, NNS) are assigned by selecting the two highest frequency subtags for each of the two remaining core tags, and reallocating the core tag probability mass between these two as in the equations above, as illustrated in Table 2.

Finally, the issue arises of what to do with the 1-to-n phrasal alignment cases shown in Figure 2 (e.g. *potatoes*/NNS → *pommes*/NNS_a *de*/NNS_b *terre*/NNS_c and *Laws*/NNS → *Les*/NNS_a *lois*/NNS_b). The potential seems to be great for function words to inherit substantial spurious probability mass via such data. However, the relatively frequent occurrence of correct 1-to-1 alignments (e.g. *The*/DT↔*Les* and *of*/IN→*de*), the diffuse nature of the noise, and the aggressive smoothing towards a single POS tag, prevent these cases from adversely affecting final function word assignments. Given the lower frequency of most content words, the potential risks of using these 1-to-n alignments are greater, but so are the benefits given that the 1-to-1 alignments tend to be both sparse and somewhat biased. Several options are under investigation for combining these two $P(t|w)$ estimators, but the simplest, and currently most effective, is to perform basic interpolation between the tag distributions estimated

from 1-to-1 alignments only and from the entire set of 1-to-n alignments (including 1-to-1) as follows:

$$P(t|w) = \lambda_2 P_{\text{1-to-1}}(t|w) + (1 - \lambda_2) P_{\text{1-to-n}}(t|w)$$

While this does indeed introduce substantial spurious tag probabilities initially, the aggressive smoothing towards the majority tag(s) described above tends to eliminate most of this noise.

4.2.2 Tag Sequence Model Estimation

The major reason for estimating the lexical priors and tag sequence model separately is that a tag sequence bigram (or even trigram) model has far fewer parameters than the lexical prior model and thus can be estimated on a very conservatively chosen set of filtered, high confidence alignment data. In contrast, the lexical prior models already suffer from sparse data problems and are negatively affected by an order-of-magnitude data reduction, even if the data is of higher quality.

The proposed model for identifying high-quality tag sequence data for training considers two different information sources for sentence filtering/weighting. The first is the final Model-3 alignment score for the sentence, indicating a multi-source measure of overall alignment confidence. The second measure more directly targets confidence in the tag sequences themselves. After the lexical prior models have been trained (as above), sentences are also tested to identify those where the directly projected tag sequence (from the automatic alignments) is closely compatible with the estimated lexical prior probabilities for each word. A pseudo-divergence weighting is computed for a sentence of length k by $\frac{1}{k}\sum_{i=1}^{k} log\hat{P}(projected\text{-}tag_i|w_i)$, penalizing words whose projected tag doesn't match the majority lexical prior.[2] Sorting and filtering/weighting by the cumulative normalized score yields a subset of training data where multiple sources essentially concur on the correct tag sequence. While the potential exists that this higher confidence data subset may be biased in the sequence phenomena it contains, the substantial noise reduction in preliminary investigations appears to be a worthwhile tradeoff. Future work will focus on differential confidence weighting of sentence fragments, and iterative (E-M) re-estimation.

[2]The exception is for function words (i.e. the majority lexical prior is not a Noun, Verb, Adjective or Adverb) located in a 1-to-n alignment sequence. Given the very high probability of these raw projections being incorrect, and their prevalence, it is expedient to attempt to correct (rather than weight/filter) these tag instances prior to the first tag-sequence-model training, by replacing their raw projection tag with the majority lexical prior for the word from 4.2.1. Doing so salvages very large quantities of otherwise accurate tag sequence data with very little introduced noise.

4.3 Evaluation of Induced Taggers

Evaluation of the tagger projection and induction algorithms is conducted on two granularities of tagset. The first tagset is at the level of core part-of-speech tags such as Verb (V), Noun (N), Pronoun (P), Adjective (J), Adverb (R), Preposition (I), Determiner (D), etc., for which English and French share remarkable compatibility.[3] The second is at the level of granularity captured in the English Penn Treebank tagset, where for example singular and plural nouns (NN and NNS) are distinguished. As previously noted, the goal of this work is not to induce potential French tagset features such as grammatical gender, mood or subtle tense distinctions that do not appear in English, but to focus on the algorithm's effectiveness at accurately transferring tagging capabilities at the granularity that is present in English (or whichever projection source language used).

For independent evaluation data, a 120K-word hand-tagged French dataset generously provided by Université de Montreal was used. However, because both this text stream and tagset had no overlap with parallel data used to train the algorithm, a simple mapping table between the tagsets was defined so that output could be compared on a compatible common denominator. An abbreviated version is shown in Table 3:[4]

Original French Tagset	English Equiv	Core Consensus Tagset
NomC-sing-⋆	NN	N
NomC-plur-⋆	NNS	N
AdjQ-⋆	JJ	J
Adve	RB	R
Prep	IN	I
Num	CD	#
ConcC	CC	C
Pron-⋆	PRP	P
Dete-⋆	DT	D
Verb-ParPas-⋆	VBN	V
Verb-ParPré	VBG	V
Verb-IndImp-⋆	VBD	V
Verb-SubImp-⋆	VBD	V
Verb-IndPas-⋆	VBD	V
Verb-IndPré-⋆	VBP	V
Verb-SubPré-⋆	VBP	V
Verb-ConPré-⋆	VB	V
Verb-InfPré-⋆	VB	V

Table 3: French-English consensus tagset map

[3]Indeed, Comrie (1990) indicates that these core POS tag distinctions tend to be almost language universal. Although some individual lexical concepts may be realized by different parts of speech in different languages, the general functional class of "noun" (for example) tends to exist in nearly all languages, and concepts which are considered to be nouns in one language also strongly tend to be realized as nouns in other languages.

[4]For compatibility with the consensus tagset, the English output tags were condensed somewhat as well, downmapping English distinctions not made in the French tagset such as comparative and superlative adjectives (JJR→JJ and JJS→JJ), the special status for 3PsingPres verbal forms (VBZ→VBP), and a separate category for modal verbs (MD→VB).

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 (auto-aligned, auto-project)	.76	.69	N/A	N/A
(b) Direct transfer (hand-aligned, auto-project)	.85	.78	N/A	N/A
(c) Standard bigram model (auto-aligned, auto-project)	.86	.82	.82	.68
(d) Noise-robust bigram induction (auto-aligned, auto-project)	.96	.93	.94	.91
(e) Standard bigram model (trained on heldout goldstandard)	.97	.96	.98	.97

Table 4: Evaluation of 5 POS tagger induction models on 2 French datasets and 2 tagsets

Because no parallel English existed for the Montreal goldstandard, to test the direct transfer models a 1000-word segment of the aligned E-F Hansard corpus was also manually labelled using this same tagset.

Table 4 shows comparative algorithm performance on each of these test sets and tagset granularities. The trend is clear in all cases: direct English-to-French tagset projection on automatically aligned data is least effective (with 76% core tagset accuracy). Yet, this problem is not entirely due to alignment errors, as direct projection from cleaned alignments only increases core tagset accuracy to 85%. Standard bigram models also perform poorly when trained on the very noisy tag projections from the auto-aligned data (86%). A 71% relative error reduction is obtained by the noise-robust induction techniques described here, with core tagset accuracy of 96% closely approaching the upper-bound 97% performance of an equivalent bigram model trained directly on 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 training on 100K words of hand-tagged training data from the identical source as the evaluation set. And while the relative difference increases to 3–6% when tested on the full English tagset granularity, this is still remarkably close for an algorithm based on entirely automatically derived, non-human-supervised data when compared with a costly hand-tagged, fully supervised learning algorithm.

4.4 Remaining Tagset Projection Issues

Despite the clear effectiveness of the algorithm in inducing a French POS tagger for both the major (core) part-of-speech distinctions and for the tag granularities realized in English, some salient French distinctions such as grammatical gender for nouns, gender and number for adjectives, and some richer verb tense/mood distinctions remain unresolved. However, most of these distinctions can be made quite straightforwardly by morphological analysis once the major part-of-speech has been resolved from context. Indeed, our parallel work has demonstrated the effectiveness of bootstrapping such morphological analyzers on monolingual text from very minimal seed exemplars (Yarowsky and Wicentowski, 2000), and via cross-lingual projection (Yarowsky, Ngai and Wicentowski, 2001). These three approaches complement each other nicely. In general, any full morphological or POS analysis clearly needs support from the other model, and the tasks should really be considered a joint effort using co-trained word-internal affixation models and context-based sequence and dependency models. However, such a union is beyond the scope of a short conference paper, and the goal of this current study is quite reasonably focused on the potential effectiveness of inducing modest-granularity POS taggers strictly from aligned bilingual corpora.

5 Noun Phrase Bracketer Induction

The second major application investigated in this paper is base noun phrase bracketing. As with POS tagging, our empirical studies suggest that BaseNP structures identified in a resource and tool-rich language such as English can be projected onto second languages via word-aligned corpora, and this noisy data can then be generalized in a robust way as a stand-alone bracketer.

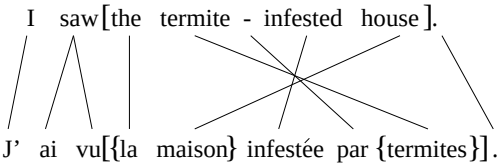

Figure 4: Example of relative noun-phrase cohesion across languages.

The essential motivation for this work is that individual noun phrases tend to cohere sequentially, and strongly resist being broken up by other sentence elements. For example, the noun-phrase in Figure 4 ("the termite-infested house"), while translated into French in very different word order, remains an unbroken sequence, resisting incursions from the temporal adverb, auxiliary verbs or other external words. This strong noun-phrase cohesion even tends to hold for relatively free word order languages such as Czech, where both informants and parallel corpus data indicate that nominal modifiers tend to re-

main in the same contiguous chunk as the nouns that they modify. As will be shown below, parallel noun phrase contiguity allows word alignments to project noun phrase bracketings as well.

5.1 BaseNP Projection Methodology

The projection process begins by automatically tagging and BaseNP bracketing the English data, using the models described in Brill (1995) and Ramshaw and Marcus (1999) respectively.

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

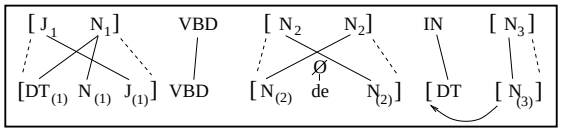

Figure 5: Standard NP projection scenarios.

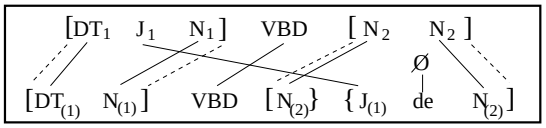

Figure 6: Problematic NP projection scenarios.

Several complicating situations exist, however: (a) Short intervening unsubscripted spans not otherwise spoken for (e.g. $[N_2\}\ IN_\varepsilon\ \{N_2]$) are incorporated into the surrounding NP. (b) Because of the high likelihood of determiners either being misaligned or aligned to null, any determiners immediately preceding a subscripted NP span are automatically incorporated into those spans, regardless of the determiner's subscript, as shown in Figure 5. (c) Finally, when two subscripted spans overlap and conflict, as in Figure 6, a likely alignment error is present and the sentence is excluded from the training data. Such a sentence is also a prime candidate for subsequent alignment repair, utilizing the BaseNP and POS models derived here.

Performance of the direct BaseNP projection procedure is detailed in Section 5.3 for both French and Chinese.

5.2 BaseNP Training Algorithm

For both model compatibility and rapid but conservative training, the Ramshaw and Marcus (1994) IOB bracketing framework and a fast transformation-base learning system (Ngai and Florian, 2001) were used. The French POS tags used as features in this process were partially based on a monolingual minimally supervised POS tagger (Cucerzan and Yarowsky, 2000) that improved slightly on the projection-based tag output described in Section 4.

As with POS tagger induction, bracketer induction is improved by focusing training on the highest quality projected data and excluding likely errorful projections. 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 hurts robustness when the stand-alone bracketers are applied to noisy tagger output.

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 of these languages, using a set of bracketing conventions that they felt were appropriate for the languages.

Performance relative to these evaluation suites was measured by exact-match bracketing precision and recall, as shown in Table 5. Studies on English, however, show that many bracketing decisions are arbitrary, and different human judges when faced with the task of bracketing several European languages typically exhibit agreement rates below 90%. Inter-judge agreement rates on our Chinese and French data were measured at 64% and 80% respectively. Similarly, the translingual projection algorithms performed here often yield perfectly reasonable bracketings that differ from the goldstandard judge by arbitrary conventions (such as whether to bracket stand-alone numbers) or by different but compatible levels of noun-phrase granularity (e.g. [DT N of N] vs. [DT N] of [N]).

Because the translingual projections are essentially unsupervised and have no data on which to mimic arbitrary conventions, an additional reasonable evaluation measure is 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 who speak the target languages further adjudicated the differences between the goldstandard and projection output as either *acceptable/compatible* or *unacceptable/incompatible*. Performance based on this measure is also included in Table 5.

Method	Exact Match			Acceptable Match		
	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 5: Performance of BaseNP induction models with precision (Pr), recall (R) and F-measure.

The large majority of these compatible divergences in bracketing convention are due to the projection algorithm's tendency to bracket possessive compounds as single NP's (e.g. [DT N *de* N]), and its tendency to bracket simple conjunctive compounds (e.g. [DT N *et* N]) also as single NPs, following the Ramshaw and Marcus convention which differed from the French and Chinese goldstandard annotator's intuitions.

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

6 Conclusion

This paper has shown that automatically word-aligned bilingual corpora can be used to induce both successful part-of-speech taggers and noun-phrase bracketers. It has further illustrated that simple direct projection of POS and NP annotations across languages is very noisy, even when the word alignments have been manually corrected. Noise-robust data filtering and modeling procedures are shown to train effectively on this low-quality data. The resulting stand-alone part-of-speech taggers and BaseNP bracketers significantly outperform the raw direct projections on which they were trained. This indicates that they have successfully distilled and modeled the signal present in the very noisy projection data, and are able to perform as respectable stand-alone monolingual tools with absolutely *no* human-supervised training data in the target language.

These results also show considerable potential for further improvement by co-training with monolingually induced morphological analyzers. The stand-alone monolingual POS taggers and bracketers induced from word-aligned data also show potential for improving their initial alignments. NP bracketings for both the source and target language can improve the IBM MT distortion model, by boosting the probabilities of word alignments consistent with cohesive NP structure, and penalizing alignments that break NP cohesion. A stand-alone POS tagger applicable to new data can be used to improve statistical MT translation models, both by supporting finer translation model granularity (e.g. *wind*/NN modeled distinctly from *wind*/VB), and by serving as a source of backoff alignment probabilities for previously unseen words. Thus tagging models induced from bilingual alignments can be used to improve these very alignments, and hence improve their own training source.

References

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.

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.

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.

B. Comrie. 1990. *All the World's Major Languages*. Oxford: Oxford University Press.

S. Cucerzan and D. Yarowsky, 2000. Language independent minimally supervised induction of lexical probabilities. In *Proceedings of ACL-2000*, Hong Kong, pp. 270-277.

I. Dagan, K. Church, and W. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Procs. of the Workshop on Very Large Corpora*, pp. 1–8.

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

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.

J. Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of ACL-93*, pp. 17–22.

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

G. Ngai and R. Florian. 2001. Transformation-based learning in the fast lane. In *Proceedings of NAACL-2001*.

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

L. Ramshaw and M. Marcus, 1999. Text chunking using transformation-based learning. In *Natural Language Processing Using Very Large Corpora*. Kluwer. pp. 157–176.

D. Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proc. ACL-94*, pp. 80–87.

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

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

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

D. Yarowsky, G. Ngai and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT-2001*, pp. 109-116.