Exploiting Hidden Meanings: Using Bilingual Text for Monolingual Annotation*

Philip Resnik

Department of Linguistics
and Institute for Advanced Computer Studies
University of Maryland
College Park, Maryland 20742 USA
resnik@umd.edu

Abstract. The last decade has taught computational linguists that high performance on broad-coverage natural language processing tasks is best obtained using supervised learning techniques, which require annotation of large quantities of training data. But annotated text is hard to obtain. Some have emphasized making the most out of limited amounts of annotation. Others have argued that we should focus on simpler learning algorithms and find ways to exploit much larger quantities of text, though those efforts have tended to focus on linguistically shallow problems. In this paper, I describe my efforts to exploit larger quantities of data while still focusing on linguistically deeper problems such as parsing and word sense disambiguation. The trick, I argue, is to take advantage of the shared meaning hidden between the lines of sentences in parallel translation.

1 The Problem of Resources

1.1 Knowledge versus Data

Success in natural language processing depends crucially on good resources. In the early days, knowledge-based approaches depended heavily on good knowledge resources — grammars, lexicons, and the like. Consider LUNAR [1], which permitted users to ask questions about moon rocks using natural language sentences. As an early question answering system, LUNAR was successful not just because of a clever formalism, but also largely because of the human effort that went into a detailed characterization of linguistic alternatives, expressed as an augmented transition network grammar and lexical entries associated with that grammar.

In the late 1980s, natural language processing began to change dramatically as the result of an influsion of ideas and techniques from the speech recognition,

* © Springer-Verlag. This paper appears in: Alexander Gelbukh (ed.), *Lecture Notes in Computer Science 2945: Computational Linguistics and Intelligent Text Processing*, Springer-Verlag, 2004, pp. 283–299.

information retrieval, and machine learning communities. Ten years ago, the "balancing act" between symbolic and statistical methods was an exciting topic for a computational linguistics workshop [2]; today it's an apt description of the entire field. Over the course of this transition, resources have remained essential, but these days what matters most is not good *knowledge* resources but rather good *data* resources. As an example, the Collins parser [3, 4] is successful not just because of a clever model, but also because of the human effort that went into annotating the Penn Treebank [5], implicitly characterizing distributions over linguistic alternatives by creating explicit linguistic representations for a large sample of text.

The impact of this shift from knowledge resources to data resources cannot be overstated. Even as the pendulum begins to swing back in the other direction, with increased attention to the potential advantages of exploiting deeper linguistic representations, the lessons of statistical methods are not being forgotten. For example, Gildea and Jurafsky [6] take a significant step toward automatic semantic interpretation by introducing a system that identifies the semantic roles, such as agent or patient, that are filled by the constituents of a sentence. They accomplish this by treating the problem as one of statistical classification, relying on a large number of sentences manually annotated with semantic role representations [7].

Annotating linguistic data with linguistic representations is undoubtedly easier than, say, writing grammars. As Sapir [8] points out, "All grammars leak" — that is, it is never possible to come up with a grammar that accurately characterizes all the data. In contrast, linguistic annotation need not be one hundred percent perfect, nor is there an expectation that it will be: instead, quality is measured by comparing the representations assigned by independent annotators to the same data, and measuring inter-annotator agreement. Variability is a given, and accurate measurement of quality given observed data is understood by all to be a part of the process. Statistics is, after all, a science of measuring uncertainty [9]. The most common use of these linguistic annotations is in probabilistic models: rather than writing rules, we estimate parameters, and by definition an estimate admits of uncertainty.

The shift from knowledge acquisition to data acquisition is not, however, free of cost. One problem that comes up again and again is the sensitivity of statistical techniques to the data on which they are trained. Good part-of-speech taggers boast accuracies within a few percent of perfect, but the published results are almost always based on experiments in which a single data set is split into training material and test material, enforcing the constraint that the training data be statistically representative of the data on which the system will be tested. In the real world, of course, this constraint can only be assumed, not enforced, and often it is not valid — taggers trained on the Wall Street Journal perform less well on data from other newswire sources, and still less well if the input text comes from another genre entirely. (To be fair, knowledge-rich approaches never did particularly well at domain-independent performance either.) This problem is gaining wider attention, and solutions are likely to come from the speech community, where automatic *adaptation* is an important topic of current research.

1.2 Acquiring Annotated Data

But of course adaptation to new kinds of data requires... data. And this is the fundamental cost in corpus-based techniques, because the techniques that perform best are all hungry for high quality annotated data. Consider parsing. Over the period from 1989 to 1992, the Penn Treebank project created skeletal parse trees for some 2.8 million words of English text, with productivity estimates (including quality control) based on annotators producing syntactic representations at an average rate of 750 words per hour [5]. The state of the art for parsing English is something on the order of 90% F-measure (harmonic mean of precision and recall) for unlabeled syntactic dependencies — and most parsers capable of producing such numbers are evaluated using the Wall Street Journal subset of the Penn Treebank, enforcing, as usual, that the training and test data are sampled from the same source. The Penn Chinese Treebank project started in the summer of 1998, and two years years later had produced an order of magnitude less data (about 100,000 words from a single newswire source), with the best parsers achieving an F-measure of 75-80%. Training good statistical parsers on the recent Chinese Treebank 3.0 release, which has 250,000 words, yields performance comparable to English (Dan Jurafsky, personal communication); but this release came out in 2003, five years after the start of the project.

If state of the art performance requires this level of annotation effort and time spent for English and Chinese, what of languages that typically receive less effort, or no effort, but suddenly become important? (For an example, see Oard et al. [10].) How can one ever hope to build annotated resources for more than a handful of the world's languages?

An interesting point of comparison is the problem of Bible translation. Here, too, is an effort to manually construct data in a large number of the world's languages, with a size that is reasonably comparable to the sizes of today's treebanks (about 800,000 words for the most-translated 66 book set, which includes both Old and New Testaments). One could argue that high quality translation is on the same order of magnitude of difficulty as detailed linguistic annotation — turnaround times for professional translation services, based on an informal survey of several Web sites, suggest a productivity estimate of around 200-300 words per hour for experienced translators. Despite a huge surge in Bible translation effort over the last century or two, and particularly in the last fifty years, today there are only approximately 400 languages for which there exist complete Bibles, and around 1000 languages for which there are New Testament translations (a difference that reflects the Bible translation community's priorities).[1] If this is the rate of progress for a task given a zealous, world-wide effort, the

[1] For current statistics, see http://www.wycliffe.org/wbt-usa/trangoal.htm. See Resnik, Olsen, and Diab [11] for a detailed discussion of the Bible as a resource for computational linguistics.

prospect for manual annotation of linguistic representations across hundreds of languages seems bleak indeed.[2]

A number of clever algorithmic solutions to the annotated data acquisition bottleneck have been and are currently being explored. One approach, active learning, pursues the idea that annotation should be done in partnership with the learning algorithm: by focusing annotation effort where it will help the learning algorithm most, rather than sampling randomly, less annotation will need to be done (see Tong [12] and references therein). Active learning is sometimes categorized as a "weakly supervised" learning approach, i.e. one in which a smaller amount of annotation data is used in combination with a large quantity unannotated data; additional avenues of research in weakly supervised learning include other general techniques such as co-training, where each of multiple learners automatically generates training examples for other learners in the set [13, 14], as well as task-specific forms of bootstrapping from small annotated subsets, e.g. in word sense disambiguation [15].

1.3 Avoiding Expensive Data Annotation

Another influential line of research, advocated recently by Eric Brill and colleagues [16, 17], suggests that instead of focusing on developing learning algorithms that make clever use of what data are available, we focus on using simpler algorithms, and obtaining a lot more data from which they can learn. This idea makes the most sense for tasks like the one explored by Banko and Brill [16], tasks for which "annotated data is essentially free" — they demonstrate that the quantity of data matters much more than the particular cleverness of the learning algorithm when disambiguating confusable word pairs such as *too* versus *to* versus *two*. (In some ways, their general point is related to Church and Mercer's [18] argument that it is better to simply collect very large quantities of corpus data than to devote effort to curating "balanced" corpora.)

For tasks such as word sense disambiguation, where the labels are not free, other novel strategies for avoiding expensive data annotation have proven useful. For example, Mihalcea and colleagues [19, 20] have explored novel ways of obtaining annotated data by taking advantage of the Web. These include both the automatic identification of new training exemplars and the elicitation of human judgments via a Web interface.

---

[2] Of course, Bible translation is not the same as translation in general, so this comparison must be treated with great caution. Wycliffe, a major organization in worldwide Bible translation, notes that translation of a New Testament can often take ten to twenty years, depending on level of participation, health of the translator (!), and other factors [http://www.wycliffe.org/wbt-usa/faq.htm]. On the other hand, Wycliffe alone has over 5100 career and short-term members, which is orders of magnitude greater than the number of people one would ever expect to find doing lingistic annotation. Readers interested in Bible translation and multilingual computing more generally should also be aware of the Summer Institute of Linguistics (SIL), which has a long history of working with lesser-known languages [http://www.sil.org].

In the remainder of this paper, I will discuss another approach to the annotated data acquisition bottleneck. It is not quite the same as doing without annotated data altogether, nor the same as restricting one's attention to shallow tasks where the data are completely free; but it is also not quite the same as performing annotation in the conventional, expensive, labor-intensive way. In Section 2, I discuss supervised versus unsupervised learning and highlight a central characteristic of the supervised paradigm. In Section 3, I argue that text in parallel translation offers this desirable property. Section 4 fleshes out the story by showing how this intuition can be applied in solving monolingual problems. Finally, Section 5 wraps up with some conclusions and discussion of work in progress.

2 Supervised Learning and Multiple Observables

I will begin the discussion with something obvious: for natural language processing tasks where both have been tried, supervised techniques generally work immensely better than unsupervised techniques. That is why there is such a demand for annotated data.

But why do supervised techniques work so much better? One answer is that by learning from annotated data, the search space over possible models is drastically reduced, in comparison to learning from unannotated data. Consider estimating the probabilities associated with a stochastic context-free grammar. Learning the parameters of the grammar — the probabilities associated with the rules — can be viewed as a search over the (very large) space of possible parameter combinations. In both the supervised case and the unsupervised case, the set of parameters is defined by the context-free structure.[3] But in the supervised case, the annotations can be viewed as providing an additional set of constraints: among sets of possible parameter values, a learning algorithm should be favoring the ones that are consistent with the observed structural annotations of the training sentences. The unsupervised learner (e.g., the Inside-Outside algorithm [21]), lacking those structural analyses, is free to find solutions that are consistent with the observed data but which are likely not to be consistent with the desired annotations. As Pereira and Schabes [22] put it, "although SCFGs provide a hierarchical model of the language, that structure is undetermined by

[3] As an aside, to repeat a point I still see too little emphasized in many discussions of statistical techniques, this is an illustration of why there is no such thing as a "purely statistical" model or method. Regardless of where the numeric parameters (the probabilities) come from, in any method, supervised or unsupervised, there is always an algebraic structure underlying the probability model; that's part of what it means for something to be a probabilistic model. Even n-gram models, which are about as "purely statistical" as one can get, embody a Markov assumption, which is equivalent to saying that their underlying algebraic structure is equivalent to a finite-state automaton.

raw text and only by chance will the inferred grammar agree with qualitative linguistic judgments of sentence structure."[4]

Another way of looking at this is in terms of what the learning algorithm gets to observe. For unsupervised algorithms, there is a set of data with representations that are observable to the algorithm, and the task is to characterize those data within the bounds imposed by an underlying model. In supervised learning paradigms, however, one finds that there are always *two* observables: a representation of the input data, and the desired output representation. Probabilistic generative models for speech recognition illustrate this. The speech stream O provides one observable, and the transcribed sequence of word tokens W is the second observable. The learning problem is to *relate* those two observables, and typically this is done by estimating a model of their joint probability $\Pr(W, O)$, given samples of the two observables paired together. In practice, speech recognition systems model the relationship by decomposing the problem into a language model $\Pr(W)$ that knows about what sorts of sequences of words get uttered and a channel model $\Pr(O|W)$ that knows how word sequences get turned into sounds.

Standard supervised classification is another illustration. Commonly the data are represented as feature vectors $\boldsymbol{x}_i$, and the learning algorithm gets to observe a sample of these paired with the classes y_i to which they belong. Imagine, for example, that $\boldsymbol{x}_i$ are collections of features that describe people, such as age, height, weight, sex, blood pressure, etc., and that y_i can be either *yes* or *no* depending on whether or not the person is considered high risk for a heart attack. Again, supervised learning can be seen as *relating* the two kinds of representations. Many methods do this by learning a separating function (e.g. a hyperplane in the feature space) that separates the $\boldsymbol{x}_i$ according to the classes to which they belong — a dividing line, so to speak, between members of the classes. Other methods, such as naive Bayes classification or maximum entropy, are better thought of as estimating the conditional probability distribution $\Pr(Y|\boldsymbol{X})$ or the joint distribution $\Pr(\boldsymbol{X}, Y)$.

This characterization of supervised learning in terms of representations is really the same as the first answer, in that the "two observables" idea describes the way in which knowledge of the desired outcomes provides a second constraint on learning, over and above the pre-determined structure of the model (stochastic CFG, noisy channel model, separating hyperplane, etc.). However, thinking in terms of observables leads to the following interesting line of thought. Supervised algorithms work well because they take advantage of two observables. Linguistic annotations, on the other hand, are unobservable representations — the standard annotation process seeks to make them observable by brute force, writing them down. Instead of giving up and summing over all the possible observables (unsupervised learning), or finding ways to perform this expensive

[4] In their paper, Pereira and Schabes demonstrate that by adding *partial* structural annotation — constituent bracketing, but without the constituent labels — the search space can still be restricted significantly, even in the absence of full parse trees.

operation on a smaller set of data (active learning), might it be possible to turn the unobservable into something that *can* be observed?

I argue that text in parallel translation provides a unique opportunity to do just that.

3 Parallel Translations and Multiple Observables

Begin with a sentence e in a single language, say, English. In addition to the observable sentence itself, we as linguists believe that there is an unobservable representation m_e of the sentence's meaning. Just characterizing a representational framework in which to express m_e is an enormously difficult problem that keeps semanticists and philosophers very busy, and it seems fair to say that actually constructing instantiations of m_e reliably for real sentences is a long way off; indeed, it's one of those problems that annotation is supposed to help us solve (e.g. [6, 7, 23]). For practical purposes, m_e is a hidden meaning.

Now consider a sentence, f, that is a translation of e. It seems safe to assume that f, too, has a hidden meaning, m_f. Let us assume, moreover, that all languages are capable of expressing the same set of meanings and use the same representational framework.[5] Then, given that e and f are translations, they express the same meaning, and $m_e = m_f$.

This fact presents an interesting opportunity. Looking monolingually, we have two observable/unobservable pairs $\langle e, m_e\rangle$ and $\langle f, m_f\rangle$. But the fact that e and f are translations of each other means that this collection of information really gives us *two observables*. The fact that they share the hidden meaning $m_e = m_f$ makes this possible, but it also means that we don't need to care about the details of that meaning representation, which is a good thing for everyone except semanticists and philosophers.

The fact that there are two observables suggests that we may be able to get some of the same advantage from parallel translations, monolingually, that can be obtained in monolingual settings where we have a single language plus its annotations: the better learning associated with supervised systems. One

[5] This assumption is potentially controversial, since there are certainly meanings that can be concisely expressed in one language that are difficult — some might even say impossible — to express in another. The concept of *Schadenfreude* is one commonly mentioned example. Another of my favorites is the American phrase *doggie bag*, which denotes an idea that can only really make sense if a meal in a restaurant is interpreted as transferring ownership of an entity rather than providing an experience. (I discovered this in France, where it appeared that the concept of bringing the dog to a restaurant meal was acceptable but the converse was not.) These objections are obviously a challenge, but at the same time, they tend to apply primarily to claims about *lexicalized concepts* in one language being translatable as lexicalized concepts in the other. Whether or not some nuances are truly untranslatable, I would argue that most of what we say to each other can be translated, if not efficiently. The concise phrase *doggie bag* can be expressed in another language as a longer noun phrase, e.g. *a package of left-over food that you bring home to eat later, or more likely put in the refrigerator and then throw away when it becomes moldy.*

language can be thought of as providing constraints on the search space for models of the other language.

This is an idea that has been explored before in a modeling setting: it is the essence of stochastic bilingual grammar formalisms. For example, Alshawi et al. [24] show that one can obtain dependency analyses for unannotated parallel text by modeling word-level alignments and syntactic dependencies together, and Wu [25] presents a bilingual grammar model that permits simultaneous learning and parsing. These approaches go part of the way toward exploiting bilingual text to perform monolingual annotation, but still suffer from a propensity to allow solutions that are justified by the data but inconsistent with linguistic intuitions — for example, Alshawi's model is perfectly comfortable identifying *collect* rather than *call* as a direct dependent of the verb in the sentence *I want to make a collect call*.[6] If the goal to is to facilitate annotating a new language with linguistically intuitive representations we hope to be able to obtain automatically, still further constraints are needed.

The further constraints come from one more assumption that tends to be true in the real world: quite frequently the language of e can be English, for which the investment in annotation has already been made, and for which, therefore, we have high accuracy algorithms for automatic annotation. The way to take full advantage of parallel text to solve monolingual problems is by using it to gain maximal leverage from the resources that have already been built.[7]

4 Using Parallel Text to Solve Monolingual Problems

4.1 Parsing

Since parsing has been a focus thus far, I begin with a discussion of my lab's experiments in automatically obtaining syntactic annotations for new languages by taking advantage of parallel text [26].

We have argued that the following assumption appears at least implicitly in almost all stochastic models that attempt to characterize the relationship between syntactic representations in two languages:[8]

---

[6] Wu [25] shows that his algorithm can be modified to take advantage of monolingual parse trees to "[facilitate] a kind of transfer of grammatical expertise in one language toward bootstrapping grammar acquisition in another", which is precisely my goal in Section 4.1. However, I am not aware of results obtained using his algorithm. In addition, what I describe is a modular approach rather than an attempt to capture everything within a single probabilistic model; see discussion in Section 5.

[7] Naturally the approach can be used in any language pair where high quality annotation exists for one of the two languages. If the approach I describe here is successful, that set will grow quickly; the greater limitation then becomes the availability of parallel text in relevant language pairs.

[8] A very recent exception is work by Jason Eisner [27] on a synchronous stochastic grammar model that permits non-isomorphic mappings between dependency trees in two languages. His model has not yet, however, been fully implemented and tested.

R	x_{Eng}	y_{Eng}	x_{Bsq}	y_{Bsq}
verb-subj	got	I	erosi	nik
verb-obj	got	gift	erosi	opari
noun-det	gift	a	opari	bat
noun-mod	brother	my	anaiari	nire

Table 1. Correspondences preserved in an English-Basque sentence pair

> **Direct Correspondence Assumption (DCA)**: Given a pair of sentences E and F that are (literal) translations of each other with syntactic structures $Tree_E$ and $Tree_F$, if nodes x_E and y_E of $Tree_E$ are aligned with nodes x_F and y_F of $Tree_F$, respectively, and if syntactic relationship $R(x_E, y_E)$ holds in $Tree_E$, then $R(x_F, y_F)$ holds in $Tree_F$.

As an example, consider this English-Basque sentence pair:

(1) a. I got a gift for my brother
    b. Nik (I) nire (MY) anaiari (BROTHER-DAT) opari (GIFT) bat (A) erosi (BUY) nion (PAST)

Table 1 shows the syntactic correspondences that are preserved under a valid word alignment between the two sentences. The correspondence is not perfect — in particular, the goal of the action is expressed as a prepositional phrase in English and via dative noun-phrase marking in Basque — but the degree of correspondence is striking, especially given that Basque and English are typologically quite distinct.

At the same time, the Direct Correspondence Assumption is just that, an assumption, and although this example is encouraging (and other encouraging examples are easy to find), it is too strong to rely upon. The distinction between a preposition and dative-marking to mark the goal phrase is but one of many examples that prove problematic. Another widespread category is the lexical expression of items in a second language that are implicit in English, such as the aspectual markers in Chinese. Since these do not appear in English there is nothing to which they can directly correspond.

One approach to this problem is to attempt to explicitly model the divergences between English and the second language [28, 29] (see also the syntactic noisy channel model proposed by Yamada and Knight [30]). Another, which I sketch here, is to accept that the correspondence may not be exactly correct, but still use the English analysis as the starting point for the second language analysis, correcting the results on the second language side. (See [26] for details.)

The algorithm for projection of syntactic dependency annotations operationalizes the DCA as directly as possible: if there is a syntactic relationship between two English words, it attempts to ensure that the same syntactic relationship also exists between their corresponding words in the second language.

Algorithm. Given word-aligned sentence pair (E, F), together with a dependency analysis of E, introduce dependencies for F according to the following cases of word-level alignment.

- **One-to-one**. If $h_E \in E$ is aligned with a unique $h_F \in F$ and m_E is aligned with a unique $m_F \in F$, then if $R(h_E, m_E)$, conclude $R(h_F, m_F)$.
- **Unaligned (E)**. If $w_E \in E$ is not aligned with any word in F, then create a new empty word $n_F \in F$ such that for any x_E aligned with a unique x_F, $R(x_E, w_E) \Rightarrow R(x_F, n_F)$ and $R(w_E, x_E) \Rightarrow R(n_F, x_F)$.
- **One-to-many**. If $w_E \in E$ is aligned with $w_{1_F}, \ldots, w_{n_F}$, then create a new empty word $m_F \in F$ such that m_F is the parent of $w_{1_F}, \ldots, w_{n_F}$ and set w_E to align to m_F instead.
- **Many-to-one**. If $w_{1_E}, \ldots, w_{n_E} \in E$ are all uniquely aligned to $w_F \in F$, then delete all alignments between $w_{i_E} (1 \leq i \leq n)$ and w_F except for the head of $w_{1_E}, \ldots, w_{n_E}$.
- **Many-to-many**. First perform the one-to-many step, then many-to-one.
- **Unaligned (F)**. Unaligned words in the second language are left out of the projected tree.

In a set of in-principle experiments on English paired with Chinese, using manually corrected parses and manually created word-level alignments, we found that direct projection of English dependencies led to poor accuracy of Chinese analyses, as measured using the precision and recall of unlabeled syntactic dependencies for a test set. However, analysis of the errors made it clear that a great many of the problems could be solved with automatic post-projection tranformations of the Chinese trees — moreover, we found that these transformations could be formulated in principled ways, taking advantage of general linguistic properties of the language, and using only a very small amount of lexically-specific information such as the identification of words in a small number of closed class categories. For example, promoting the initial word in a multi-word constituent to be the head, making the remaining words its dependents, had a dramatic effect on accuracy, understandably, because Chinese is for the most part a head-initial language. Other refinements included modifying the head-promotion rule to pay attention to parts of speech (like English, the Chinese nominal system is head-final) and additional operations such as attaching an aspectual marker to the preceding verb. Within two or three person-days of language-specific effort, the quality of the projected trees, based on ideal parses and alignments, had improved dramatically, as illustrated in Table 2.

These results demonstrated that, in principle, reasonably high quality trees for Chinese could be obtained via projection of English analyses followed by post-projection transformations. (Recall that automatic statistical parsers trained on the Penn Chinese Treebank, after two years of manually intensive treebank construction, typically obtained an F-measure of only 75-80%.) We hoped that by trading off somewhat reduced quality for quantity — creating a treebank via projection with an order of magnitude more Chinese sentences — it would be

Method	Precision	Recall	F-measure
Direct	34.5	42.5	38.1
Head-initial	59.4	59.4	59.4
Other rules	68.0	66.6	67.3

Table 2. Quality of projected dependency analyses for Chinese, using clean English parses and correct alignments (%)

Method	Training corpus	Sentence pairs	Parser accuracy
Baseline	-		34
Statistical parser	UN/FBIS/Bible	98,000	67
Statistical parser	UN/FBIS/Bible (filtered)	20,000	72
Commercial parser	-	-	69

Table 3. Quality of parser performance for Spanish, training on a treebank that was obtained *fully automatically* via projection of annotations from English followed by a small set of post-projection transformations (%)

possible to make up for the fact that, even with ideal English parses and idea word alignments, the projected trees are fairly noisy.

However, it turned out that English-Chinese projection did not work as well using automatically obtained parses and word alignments. On further analysis, varying the quality of both the parser and the word alignments, it became apparent that the barrier to high quality automatic projection is not the automatic parsing on the English side, but rather the quality that could be obtained in automatic word alignment between English and Chinese.

In order to confirm that this was indeed the case, we followed up our English-Chinese experiment with an experiment in automatic projected treebanking using English-Spanish as a language pair, since English and Spanish are much easier to automatically align. We followed precisely the same paradigm, including a several person-day effort to develop Spanish-specific post-projection transformations. Table 3 shows the results.

The table provides a comparison against a simple baseline in which every word is considered a dependent of the previous word. It also shows the benefit of automatically filtering the projected trees in order to remove cases that are likely to hurt rather than help in parser training. Filtering criteria included, for example, English-Spanish alignments in which too many words were unaligned, alignments in which there was an n-to-1 alignment for too large an n, and cases where the projected tree contains too many crossing dependencies. (All thresholds, etc., were manually tuned on development data prior to test, of course.) The key result here is the performance of the statistical parser in comparison to a state-of-the-art, rule-based commercial parser for Spanish. Without manual annotation of any Spanish training data, and with only a few days of linguistically informed effort (compared to months or years of grammar writing), this approach yields a Spanish parser with state of the art performance.

4.2 Word Sense Disambiguation

As is the case for parsing, there is a dramatic difference in the performance of supervised versus unsupervised learners in word sense disambiguation (WSD). Unlike parsing, the state of the art even for supervised annotation of word senses is not high enough to inspire confidence that the sense tags should be projected to a second language. The question arises, therefore, as to whether the information present in parallel texts can be used to help improve *English* word sense tagging — and secondarily, perhaps, to bootstrap WSD in the second language, as well. Mona Diab explored this question in her doctoral dissertation [44]; here I sketch the approach and results reported by Diab and Resnik [31].

It has long been observed that translation into distinct foreign language words often makes a sense distinction evident [32–39]. The oft-cited financial and riverbank senses of *bank*, for example, are realized in French as *banque* and *rive*, showing how the two meanings, hidden behind the same word, can be made observable by noting a lexical correspondence with another language. If the word *bank* occurs in an English sentence, and it corresponds to *banque* in the French translation, then clearly the financial sense is intended.

In considering how parallel translation might provide a useful set of constraints for learning to disambigate word senses, it is interesting to observe that the converse also holds: when two distinct English words can be translated as the *same* word in a second language, it often indicates that the two are being used in senses that share some element of meaning. For example, *bank* may be ambiguous, as just noted, and so might the word *shore* (it can refer both to a shoreline and to a piece of wood used as a brace or support). But both *bank* and *shore* can be translated into French as *rive*, and this fact suggests that the two senses corresponding to that translation have something in common — some semantic quality we might call *rive*-ness, so to speak.

To the extent that it is reliable, this additional information available from the second language translation can be exploited in order to narrow in on English word senses, even in the absence of sense-tagged English training data. Algorithmically, the process requires treating a second-language word like *rive* as an anchor point that delineates a set of putatively-related English words. For example, analysis of a word-aligned parallel corpus might produce the information that French *rive* shows up with the set of English words *bank*, *shore*, and *shoreline*.

This information can be used to establish preferences among senses in an English sense inventory, so long as that sense inventory provides some way to measure the similarity of (or distance between) word senses. Intuitively, in those cases where *bank* is translated as *rive*, one can assume as a working hypothesis that it is being used in the sense that is most similar to the senses of *shore* and *shoreline* when *they* can be translated as *rive*. Operationally, it is straightforward to design an algorithm that takes a set of related words such as {*bank*, *shore*, *shoreline*}, and reinforces the words' different senses differentially depending on their similarity to the senses of other words in the set [40].

One of the problems with using this idea for sense disambiguation is evaluation: there are many parallel corpora, and there are a number of relatively standard WSD test sets, but it is virtually impossible to find a test set that exists in parallel translation with a second language, in order to evaluate the extent to which the second language could help in disambiguation. Rather than manually sense-annotating a parallel corpus, which would have been quite expensive, we instead opted to use a carefully sense-tagged test set for English and to use machine translation to produce a pseudo-parallel corpus.

Across a range of conditions (varying the languages, the machine translation system used to provide pseudo-parallel text, and experimenting with combining information across languages), the approach performed consistently well on the SENSEVAL-2 test data, achieving precision in the vicinity of 60% and with recall around 54% for disambiguation of nouns in the "All Words" task. (See Diab [44] for details and additional performance improvements, as well as for experiments using human-translated data and other extensions.) This compares quite favorably with the performance of the other unsupervised SENSEVAL-2 systems, and in fact it places this approach on par with a number of the supervised systems.

Interestingly, these results were obtained not only without any manually sense tagged training data, but also without taking advantage of any monolingual context whatsoever. There is, therefore, good reason to believe that further improvements can be obtained by combining the cross-language approach with the better unsupervised methods for sense classification.

5 Conclusions

Supervised techniques have revolutionized natural language processing; as a result, data annotation is its greatest bottleneck. In this paper I have suggested a strategy that complements other approaches to reducing the annotation burden: using the constraints provided by parallel translation to create noisy annotations of large quantities of text, rather than focusing on high quality annotation for smaller data sets. Because the constraints implicit in parallel translation are still not a guarantee that the resulting annotations will look like what is needed, the strategy gains leverage from the existence of high quality broad-coverage annotation tools for English.

The research described here is intimately connected with the work of David Yarowsky and students [41–43]. Their ground-breaking work on the projection of shallower representations across parallel text in order to bootstrap language analyzers (including part of speech taggers, morphological analyzers, noun phrase bracketers, and named entity taggers) is complemented by the work I have described here, which focuses on linguistically deeper representations such as syntactic dependencies and word senses. These are two facets of a joint research project focused both on improving statistical machine translation with richer linguistic features and on rapidly bootstrapping monolingual analyzers for new languages. We refer to the process of creating analyzers via projection of representations from English as the annotation-projection-training approach, or just as "annotation projection."

I have already briefly discussed several similarly motivated lines of work that attempt to account for both rich monolingual representations and cross-language relationships in a single model, e.g. Alshawi et al. [24], Eisner [27], and Wu [25]. Such approaches certainly are capable of taking advantage of constraints from high quality English annotations in their modeling. I have tended to avoid this style of modeling in favor of actually projecting explicit representations, however, for a number of reasons. One is that the barriers to entry are much lower when it is possible to work with modular software components; one need not implement an end-to-end probabilistic framework from scratch, with all that entails, particularly the complexities of parameter estimation and search. As a closely related issue, adopting a modular architecture rather than a unified model makes it easy to swap in and out components such as English analyzers, word alignment methods, and supervised learning algorithms.[9] Finally, and most important, I am a great believer in taking advantage of independently motivated development activity, in this case the rapid progress being made in supervised learning methods for natural language processing.

I am currently pursuing several continuations of the current research. In parsing, I am exploring the possibility that post-projection linguistic transformations can be made unnecessary. The key observation is that, given a small treebank in a language of interest, it is possible to train a parser that captures some of the most important language-specific phenomena (e.g. aspectual marker attachment in Chinese), even if its training data are too sparse for it to take advantage of, for example, lexical conditioning of syntactic rules. At the same time, the annotation projection approach can produce large numbers of trees that capture a great deal of valid knowledge carried over from English, but which leave a parser ignorant of language-specific constraints. Students and I are exploring an iterative approach to expanding the small, high quality monolingual treebank using parallel text. Given a large parallel corpus of sentence pairs $\langle e, f\rangle$, the idea at iteration i is to parse a new sentence f with the currently trained small-treebank-parser π_i, producing n-best "monolingually informed" trees, $\{T_\pi^1, T_\pi^2, \ldots, T_\pi^n\}$. Simultaneously, we project syntactic annotations from e to produce m-best "English-informed" projected trees for f, $\{T_E^1, T_E^2, \ldots, T_E^m\}$. By combining knowledge from the English-informed and monolingually-informed tree sets for f, it should be possible to select the "best" of the automatically obtained T_π trees for addition to the training set at iteration $i+1$, as judged not only on its monolingual confidence score (a variant of self-training), but also on the extent to which the relationships it encodes are consistent with the English analysis. It should also be possible to create a "consensus" analysis that com-

[9] Recent developments by Jason Eisner may affect these considerations; his Dyna programming language is designed to facilitate exploration of alternative models by automating many of the implementation details. See http://www.cs.jhu.edu/~jason/dyna/.

bines the most confidently assigned dependencies from the T_π and T_E analyses. For example, in the English-Chinese case, the T_E trees may be able to confidently identify the verb-subject and verb-object relations, and the T_π trees may be able to provide confident attachment points for aspectual markers.

In word sense disambiguation, I am exploring a similar combination of unsupervised and supervised methods in order to identify larger numbers of training samples. In this case, a large parallel sample will be used as the basis for bilingual unsupervised sense tagging (in the style of Diab [31, 44]). A small portion of held-out annotated data will be used to develop a classifier for confidence estimation, in order to identify automatically sense tagged items above a confidence threshold. These high-confidence items will be added to the training sample, and the process will continue iteratively.

Ultimately, the goals of this work are three-fold. First, as has been emphasized in this paper, it is to be hoped that by gaining leverage from parallel translation, it will be possible to develop linguistically deeper monolingual analyzers for a wider range of the world's languages. Second, there is, of course, a close connection between this work and the development of more sophisticated statistical models for machine translation — it is widely agreed among statistical MT researchers that many of the shortcomings of current technology have their source in a less than adequate treatment of syntactic behavior, which is leading to increased exploration of a syntactically more sophisticated set of statistical models. And finally, there is a goal that is connected with these technological goals only indirectly: it is to be hoped that by exploring the way different languages express the same hidden meaning, facilitated by the tools and models of statistical NLP, we may come closer to understanding the nature of language itself.

Acknowledgments

Work described in this paper was done in collaboration with Mona Diab, Rebecca Hwa, Okan Kolak, and Amy Weinberg; please see works cited for details. The author gratefully acknowledges the support of Department of Defense contract RD-02-5700, DARPA/ITO Cooperative Agreement N660010028910, and ONR MURI Contract FCPO.810548265. This paper was written while the author was on sabbatical at Johns Hopkins University; discussion of this research with colleagues and students there was much appreciated. Any errors are solely the responsibility of the author or the author's dog.

References

1. Woods, W.A., Kaplan, R.: The lunar sciences natural language information system. Technical Report 2265, Bolt, Beranek, and Newman, Cambridge, MA (1971)
2. Klavans, J., Resnik, P., eds.: The Balancing Act: Combining Symbolic and Statistical Approaches to Language. MIT Press (1996) http://mitpress.mit.edu/book-home.tcl?isbn=0262611228.

3. Collins, M.: Three generative, lexicalised models for statistical parsing. In: Proceedings of the 35th Annual Meeting of the ACL, Madrid (1997)
4. Collins, M.: Head-driven statistical models for natural language parsing. Computational Linguistics (to appear)
5. Marcus, M.P., Santorini, B., Marcinkiewicz, M.: Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics **19** (1993) 313–330
6. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics **28** (2002) 245–288
7. Fillmore, C.J., Johnson, C.R., Petruck, M.R.: Background to FrameNet. International Journal of Lexicography **16** (2003) 235–250
8. Sapir, E.: Language: An Introduction to the Study of Speech. New York: Harcourt, Brace (1921) www.bartleby.com/186/.
9. Stigler, S.M.: The History of Statistics. Harvard University Press (1986)
10. Oard, D.W., Doermann, D., Dorr, B., He, D., Resnik, P., Weinberg, A., Byrne, W., Khudanpur, S., Yarowsky, D., Leuski, A., Koehn, P., Knight, K.: Desperately seeking Cebuano. In: Proceedings of the HLT-NAACL Conference. (2003) 76–78 Late Breaking Results.
11. Resnik, P., Olsen, M.B., Diab, M.: The Bible as a parallel corpus: Annotating the 'Book of 2000 Tongues'. Computers and the Humanities **33** (1999) 129–153
12. Tong, S.: Active Learning: Theory and Applications. PhD thesis, Stanford University (2001) http://www.robotics.stanford.edu/∼stong/papers/tong_thesis.ps.gz.
13. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers. (1998)
14. Sarkar, A.: Applying cotraining methods to statistical parsing. In: Proceedings of the Second North American Conference on Computational Linguistics (NAACL 2001). (2001)
15. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, Association for Computational Linguistics (1995) 189–196
16. Banko, M., Brill, E.: Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. In: Human Language Technology Conference (HLT). (2001)
17. Brill, E., Lin, J., Banko, M., Dumais, S., Ng, A.: Data-intensive question answering. In: Proceedings of TREC 2001. (2001)
18. Church, K.W., Mercer, R.: Introduction to the special issue on computational linguistics using large corpora. Computational Linguistics **19** (1993) 1–24
19. Chklovski, T., Mihalcea, R.: Building a sense tagged corpus with open mind word expert. In: Proceedings of the ACL 2002 Workshop on 'Word Sense Disambiguation: Recent Successes and Future Directions'. (2002)
20. Mihalcea, R.: Bootstrapping large sense tagged corpora. In: Proceedings of the 3rd International Conference on Languages Resources and Evaluations (LREC 2002). (2002)
21. Baker, J.: Trainable grammars for speech recognition. In: Proceedings of the Spring Conference of the Acoustical Society of America, Boston, MA (1979) 547–550
22. Pereira, F., Schabes, Y.: Inside-outside reestimation from partially bracketed corpora. Proceedings of the February 1992 DARPA Speech and Natural Language Workshop (1992) 122–127

23. Kingsbury, P., Palmer, M., Marcus, M.: Adding semantic annotation to the Penn TreeBank. In: Proceedings of the Human Language Technology Conference (HLT'02). (2002)
24. Alshawi, H., Srinivas, B., Douglas, S.: Learning dependency translation models as collections of finite state head transducers. Computational Linguistics **26** (2000)
25. Wu, D.: Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics (**23**) 377–403
26. Hwa, R., Resnik, P., Weinberg, A., Kolak, O.: Evaluating translational correspondence using annotation projection. In: 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia (2002)
27. Eisner, J.: Learning non-isomorphic tree mappings for machine translation. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics (companion volume). (2003)
28. Dorr, B.J., Pearl, L., Hwa, R., Habash, N.: DUSTer: A method for unraveling cross-language divergences for statistical word-level alignment. In: AMTA. (2002) 31–43
29. Drabek, E., Yarowsky, D.: personal communication.
30. Yamada, K., Knight, K.: A syntax-based statistical translation model. (In: Proceedings of the 39th Meeting of the Association for Computational Linguistics (ACL 2001)) 523–530
31. Diab, M., Resnik, P.: An unsupervised method for word sense tagging using parallel corpora. In: 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia (2002)
32. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: A statistical approach to sense disambiguation in machine translation. In: Proc. of the Speech and Natural Language Workshop, Pacific Grove, CA (1991) 146–151
33. Dagan, I.: Lexical disambiguation: sources of information and their statistical realization. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. (1991) Berkeley, California.
34. Dagan, I., Itai, A.: Word sense disambiguation using a second language monolingual corpus (1994)
35. Dyvik, H.: Translations as semantic mirrors. In: Proceedings of Workshop W13: Multilinguality in the lexicon II, Brighton, UK, The 13th biennial European Conference on Artificial Intelligence ECAI 98 (1998) 24–44
36. Ide, N.: Cross-lingual sense determination: Can it work? Computers and the Humanities: Special issue on SENSEVAL **34** (2000) 223–234
37. Ide, N., Erjavec, T., Tufis, D.: Automatic sense tagging using parallel corpora. In: Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium. (2001) 83–89
38. Ide, N., Erjavec, T., Tufis, D.: Sense discrimination with parallel corpora. In: Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. (2002) 54–60
39. Resnik, P., Yarowsky, D.: Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. Natural Language Engineering **5** (1999) 113–133
40. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research (JAIR) **11** (1999) 95–130
41. Yarowsky, D., Ngai, G.: Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In: Proceedings of NAACL-2001. (2001) 200–207

42. Yarowsky, D., Wicentowski, R.: Minimally supervised morphological analysis by multimodal alignment. In: Proceedings of ACL-2000. (2000) 207–216
43. Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. (In: Proceedings of the First International Conference on Human Language Technology Research (HLT 2001),)
44. Diab, M.: Word Sense Disambiguation within a Multilingual Framework. PhD thesis, University of Maryland (2003)