Cross Lingual Syntax Projection for Resource-Poor Languages

Vamshi Ambati
Language Technologies Institute,
Carnegie Mellon University

Wei Chen
Language Technologies Institute,
Carnegie Mellon University

1. Introduction

Over the past few decades, supervised learning in structured spaces has been quite successful in syntactic analysis problems in natural language processing. These learning techniques exploit large amounts of annotated data to learn models that can perform linguistic analysis on unseen data. Acquiring such supervised linguistic annotations for a language is important for natural language processing and it usually involves significant human efforts. The quantities of the annotated data are far from being sufficient for the majority of the languages. Languages like English have been well supported in the linguistics community, and therefore there is a wealth of language analysis tools for them. We also have large amounts of annotated data available. This makes English a resource-rich language and attractive for computational linguists to work on. There are only a few more languages in the world that enjoy the status of a resource-rich language. Many other languages either do not have analysis tools or do not have annotated data from which state-of-the-art tools can be induced. This makes these languages resource-poor both in terms of data and tools. Even after 50 years of notable contributions made in the area of computational linguistics, we are still far from being able to deal with many other languages.

The advent of the World Wide Web and the advances in the digital media world have helped the language community immensely. We now see a lot of data on the internet and a lot of parallel data for various languages pairs. There are also known techniques for harvesting parallel texts from the World Wide Web (Resnik and Smith 2003). A pair of texts is parallel when a document in one language, often the source language, has an identified mapping with another document in the second language, called the target language, and one is an equivalent translation of the other.

The availability of parallel data has opened various research ideas for the creation of multilingual applications, in particular for the resource-poor languages. One approach is to project syntactic annotations and structures from the resource-rich source language to the target language which is resource-poor. This is often called "projection of annotation" or simply "syntax projection". The goal of syntax projection is to induce multilingual text analysis tools automatically for a target language. This problem can be complicated because of the differences in syntactic structures between the source and target languages. Usually, the projection is not merely a one-to-one mapping. Rather, syntactic relations can also be one-to-many, many-to-one, many-to-many or even unmapped. Annotated data that we get from direct projection using parallel corpora contains errors. Thus, training accurate stochastic text analyzers from noisy data becomes a challenging task. Many efforts have been developed and put into practice in the last 10 years to solve the challenges faced by syntax projection (Yarowsky, Ngai, and Wicentowski 2001; Hwa

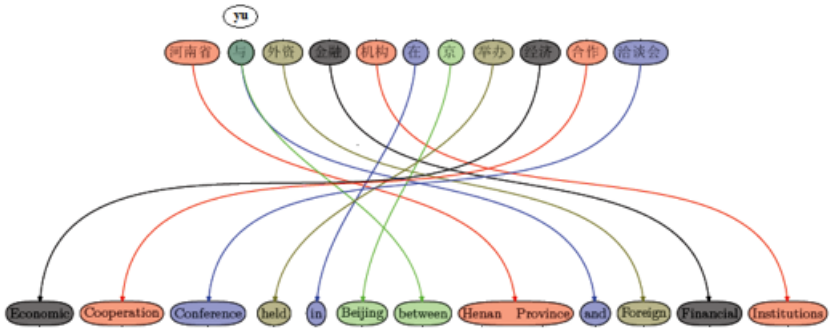

Figure 1
Parallel sentence example

et al. 2005; Resnik 2004). The two main challenges faced are word-alignment error and the syntactic divergences of the two languages. In our survey we discuss the paradigms of syntax projection and the application of these approaches to various kinds of syntactic annotations. We also discuss how these various techniques have dealt with the main challenges of syntax projection and have applied it successfully to larger problems of natural language processing like machine translation (Ahmed and Hanneman 2006).

The rest of the report is organized as follows. In the section 2 we first describe and formalize the task of syntax projection and motivate the main challenges of syntax projection. In Section 3, we discuss various kinds of syntactic annotations in natural language and categorize them into relevant structured spaces for syntax projection. In Section 4, we first discuss grammar-based methods for syntax projection. In Section 5, we then survey heuristic-based approaches for syntax projection that most often require word correspondences between the two languages as a prerequisite. Section 6 discusses the common methods of evaluation for the syntax projection task. Section 7 concludes the report by broadly pointing to the application of these techniques in other areas of natural language processing.

2. Syntax Projection across Languages

In this section, we describe basic concepts in syntax projection and use an example to show some challenges in this task. Given a parallel corpus $D(S,T)$ and an annotation model A_s for the source language S, the task of syntax projection is to infer the annotation model A_t for the target language T, where the parallel corpus $D(S,T)$ consists of texts in the source language S and their translations in the target language T, and the annotation model A is used for annotating raw texts in one language.

Parallel text data contains three kinds of information: sentences in a source language, their translation sentences in a target language, and the alignment information between the sentence pairs. Alignment information is usually represented as a list of ordered pairs of indices of the words in the sentences. Because of the language diversity, translations of one sentence in multiple languages may vary a lot in their word order. Thus, alignment information is very helpful in modeling syntax projection across languages.

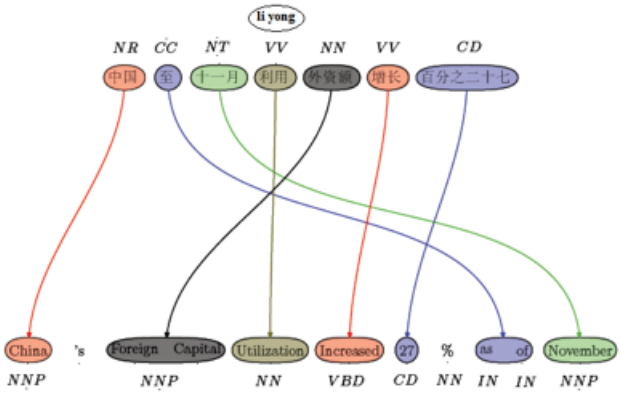

Figure 2
A problem of direct projection in part-of-speech tagging

Figure 1 shows an example of an English-Chinese parallel sentence. In this example, the alignment between the word pairs is visualized as links between the words in the two languages. Notice that each word in the example is mapped to at least one word in the other language. This kind of full mapping does not always happen in real corpora. In reality, each word can map to single, multiple, or zero words in the other language. This phenomenon stands as a challenge to syntax projection. Take part-of-speech (POS) tagging as an example. Imagine that we only have exact one-to-one mappings in the parallel text; then directly projecting parts of speech to the target language seems to solve the problem. However, in English-Chinese parallel text, most sentence pairs do not have this property. In Figure 1 "between .. and .. " is translated into one single Chinese word "yu" [1]. Also, we know that "between" and "and" do not share the same part-of-speech ("between" is a preposition, while "and" is a conjunction). This causes a difficulty for deciding the part-of-speech for the Chinese word "yu". In fact, even for one-to-one mapping, direct projection may not give the right answer. For example, in Figure 2, the fourth Chinese word "li-yong" is a verb in the Chinese sentence, but its English translation "utilization" is a noun. Another observation is that the POS tagsets for Chinese and English may be different. In other words, using the English POS tagset for projection into Chinese may cause a problem. We will go back to these issues in more detail in our discussion of methods in section 5.

We should now realize that although alignment information is helpful, it is not sufficient for syntax projection. Parallel corpora usually come from human translations, and good translations are not word-to-word mappings. One word or phrase in a language may be translated in a very flexible way, since the goal of translation is to preserve the meaning, rather than syntactic structure. We have presented POS tagging as an example to show some challenges in syntax projection. We should also notice that these problems are by no means exhaustive. Specific problems may occur in specific applications. Usually, different methods and assumptions used for syntax projection come from careful observations of particular tasks and language pairs.

1 We use Pinyin for the transliteration of Chinese

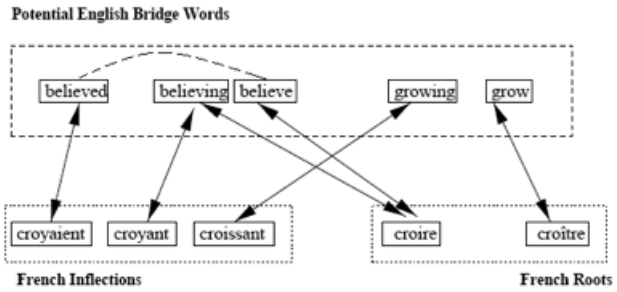

Figure 3
Morphology projection example: Adapted from Yarowsky and Ngai 2001: Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora

3. Syntactic Structure Spaces

The approaches to syntax projection are directly influenced by the kind of syntactic annotations that we intend to project. In this section, we classify syntactic structures into three categories: individual lexicons, flat sequential structures, and hierarchical structures. We use examples to discuss challenges for syntax projection with respect to each category.

3.1 Individual Lexical Annotations

Individual lexicon annotations include dictionary annotation, morphological analysis, and other lexicon annotations that do not involve contextual information. By this, we mean that the output of the annotation model is individual lexicons, rather than sentences or other sequential structure. However, recent methods for learning such annotations might make use of context (Probst 2003). We use morphology induction as an example to illustrate challenges in individual lexical annotation projection.

Research in morphology is concerned with the way that words are built up from morphemes, the smallest units of meaning. Morphological rules can vary a lot from language to language. Some languages are highly inflective, such as Hebrew and Czech, while some others do not have morphology at all, like Chinese. The major problem in morphology induction comes from the irregular cases, where an induction does not follow the basic rules or the root form. Yarowsky, Ngai, and Wicentowski (2001) show that a bilingual parallel corpora can be very helpful when analyzing morphology induction. Figure 3 shows an example, where the French word "croyant" is associated with its root "croire" through the English bridge word "believing". Notice that the links with arrows are actual alignments existing in the parallel corpus. The problem for this approach is that such direct mappings are usually rare, leaving a large amount of root and its inflected forms unresolved. For example, in the same Figure, another French word "croyaient" cannot be linked to its root form "croire" because there is no alignment between "believed" and "croire". Fortunately, the gap can be filled by the relationship of "believed" and "believe" on English side. Through this way, "croyaient" can be successfully associated with its root "croire".

The key idea here is that individual lexical annotations usually may not be projected through direct mapping because of missing links in parallel corpora. There are several

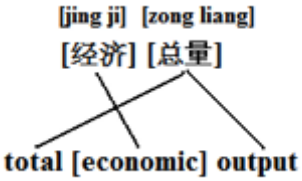

Figure 4
Problem of noun-phrase bracketing due to non-trivial mapping

reasons for this problem. Firstly, we may not have enough data to include all possible alignments. Secondly, even if alignments are complete, we may still have missing links. In the English-French example, "croyaient" is not linked to "believe" because "croyaient" is a past tense verb, and it may never map to an infinitive form. Dictionary annotation has similar problems. For example, tagging number information on adjectives in some languages faces the problem that the source language used in transfer does not have this information (Probst 2003). Thus, the gap needs to be filled by information provided by context. Usually, the English nouns closest in distance in the sentence are chosen to tag the number information of the adjectives. We will delay details for the models used in these approaches to Section 5.

3.2 Flat Sequential structure annotations

Flat sequential structure annotations include POS tagging, named-entity tagging, base noun-phrase bracketing, and other sequential annotations without hierarchical structures. Since sequential structure projection involves contextual information, the problem of parallel text (translating meaning rather than syntax) mentioned in the previous section comes back. Nouns, verbs, adjectives, and adverbs are usually translated directly to convey the full meaning, so these words are often used for experiments on POS projection. Others, like prepositions, usually do not correspond one-to-one or have an equivalent in translation and so are likely to be excluded. From the previous examples shown in Figure 2, we see that one major challenge for POS projection comes from not having one-to-one mappings. This is a common difficulty in flat sequential annotation projection. Figure 4 shows an example of noun-phrase bracketing, where the second Chinese word "zong liang" maps to two English words, and the two are separated by another word "economic".

Research in named-entity recognition is slightly different from the other flat sequential annotations. The goal is not to project named entities. Rather, the goal is to recognize named entities for one language with the help of parallel texts. Klementiev and Roth (2006) propose a method for named-entity recognition in Russian with temporally aligned English-Russian parallel texts. With the knowledge of named entities in English, a measurement of similarities between English and Russian words, and other linguistics observations, named-entity discovery can be resolved in a more robust way compared with monolingual methods.

3.3 Hierarchical Structure Annotations

Hierarchical structure annotations include dependency trees, phrase structure trees, and semantic role labeling. Hierarchical structure projections have the same problem with flat sequential structure projections which comes from various kinds of mappings of words. But it can be even more complicated. For example, direct mapping of dependency structures from English to Chinese would result in non-projective dependency trees (Ryan T. McDonald and Hajic 2005). Besides, it is hard to decide the dependency relations for unaligned words. Further, direct mapping of phrase structures would result in illegal phrase structure trees, where one constituent may cross other constituents. Figure 5 illustrates a projection from an English phrase structure tree to a Chinese tree. The yield of the chinese tree contains the English translation of the Chinese words. We could see from the surface that the structures of the trees on two sides are quite different. Because of these problems, researchers in tree structure projections usually make specific assumptions and lists of rules based on observations of particular language pairs to simplify the problem (Xi and Hwa 2005). Because of the special difficulty in projecting tree structures, a post-projection transformation phase is usually involved to correct and filter the output. This requires considerable knowledge of the target language.

We also include semantic role labeling in hierarchical structure annotation because it involves relations and their arguments, which can also be other relations. Thus it is a hierarchical structure. The problem in semantic role labeling is similar to tree structure annotations, although the approaches to these problems can be quite different. For robustness, non-content words are usually dropped in experiments, as is mentioned in the example of POS tagging. Figure 6 shows an example of semantic role label projection from English to Chinese. In this example, the relationship (leadership) and its arguments (Taiwan and Authorities on Taiwan) are projected to Chinese through direct mapping. Very similar to tree structure projections, semantic role labeling also requires post-processing for acceptable accuracy.

3.4 Summary

We have introduced three categories of syntactic structures and the different challenges for syntax projection. Generally, flat sequential annotation projections are more complex than individual lexical annotations, since they involve more context relations. And hierarchical structure projections are more complex than flat structures, since more constraints are involved in, and more variations can happen. Because of these issues, hierarchical structure projections usually require an additional step called postprocessing to clean the noisy outputs. Understanding challenges in different syntactic structures will help us better comprehend the approaches used in syntax projection, which will be discussed in the following sections.

4. Grammar-Based Approaches to Syntax Projection

In this section we summarize the approaches of syntax projection that implicitly or explicitly use the grammatical structure of the target language into which the projection is done. The target grammar could be incorporated into the process of projection in several different ways. We first discuss approaches that use a synchronous grammar to perform the parsing of both languages in lock-step, thereby creating syntactic structures for the target side. Although many such formalisms that model parallel sentences exist, in this section we discuss some of the formalisms that are specifically applied to the task

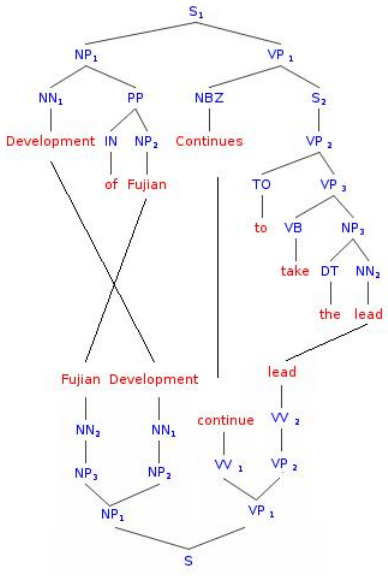

Figure 5
Parallel English Chinese parse trees and phrase structure projection

of syntax projection. We will also look into other approaches that treat the task of syntax projection as the problem of finding the optimal target syntax structure, given the source grammar, linguistic knowledge of the target language and the correspondences for the two languages.

4.1 Inversion Transduction Grammars

Wu (1997) proposes a novel extension to transduction grammars of the finite state family to handle bilingual language modeling and parsing called the Inversion Transduction Grammars (ITG). ITGs relax the monotonicity constraint imposed by the transduction grammars. While transduction grammars only have a straight orientation on its productions in both the input and output streams, ITGs allow for an inverted orientation. This makes ITGs quite useful for natural language processing tasks like bilingual parsing where both the languages are syntacticly divergent and the grammars should allow for the inversion of the constituents. A typical ITG, expressed in a 2-normal form, looks as shown below. Rules are of the form $A \rightarrow x \ / \ y$, where A is the non-terminal that generates two symbols x and y in two simultaneous streams, very often referred to as input and output streams. The rules also allow for essentially not producing any symbol in either of the streams. Rules that generate non-terminals are usually enclosed in square brackets and indicate that the same sequence is also produced in the second stream. The last rule in the grammar, where the production is enclosed in angular brackets, is the interesting rule which allows for the inversion, where B and C are inverted in the output stream.

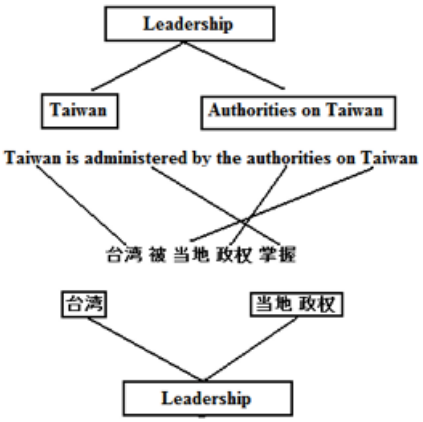

Figure 6
Semantic role labeling projection from English to Chinese

$$S \rightarrow \epsilon \ / \ \epsilon$$
$$A \rightarrow x \ / \ \epsilon$$
$$A \rightarrow \epsilon \ / \ y$$
$$A \rightarrow [x \ y \ ]$$
$$A \rightarrow [B \ C \ ]$$
$$A \rightarrow <B \ C >$$

Given a pair of sentences, parsing using the ITGs means to identify the matching constituents on both sides, which are not necessarily linguistically motivated constituents. (Wu 1997) also discusses a stochastic version of the ITGs called stochastic inversion transduction grammars (SITGs) where every production is associated with a probability. This now models a more realistic scenario of parsing a pair of sentences and identifying bracketing on both sides. Although the primary motivation of SITGs was bilingual sentence modeling, Wu (1997) also discusses the application of SITGs to a scenario where one side of the parallel corpus is a well studied language like English that has a parse tree available. The SITGs is then applied to optimize the bilingual parsing in conjunction with the available source-side syntax. One drawback of the approach is the excessive reliability on the word alignments in the formalism, which creates a problem when there is not enough data to train on or if the languages are drastically different in word orders. This is however a very novel piece of work that has motivated much of other work in syntax based statistical translation systems (Ahmed and Hanneman 2006).

4.2 Synchronous Grammar Models

A number of synchronous grammar formalisms have been proposed in the past decade for the task of bilingual parsing. Shieber and Schabes (1990) describes a synchronous tree adjoining grammar, while Melamed (2003) proposes a more general version of bilingual grammars called multi text grammars and also discusses algorithms for pars-

ing them. While many of these grammars are directly applicable in the context of machine translation or bilingual parsing, combining them with a word correspondence model and inferring them in the context of resource-poor languages makes them more interesting for the task of syntax projection.

The previous grammar formalisms are limited in certain ways; for example, the SITGs (Wu 1997) assumes that only the leaf nodes or the terminals can produce NULL values, but other non-terminal nodes can produce equivalent non-terminals in the second language in either a monotonic or non-monotonic manner. Also there are implicit assumptions of the source and target syntax structures having a plausible mapping between the nodes, as well as mapping at the word level alignments of the sentences. Smith and Smith (2004) relax some of the assumptions by using any amount of information provided as a probabilistic n-best outputs of the individual models. They propose a unified log-linear model to combine an English parser, the word alignment model, and a Korean PCFG parser trained from a small number of Korean parse trees. The basic grammar formalism and idea of biparsing is similar to a multitext grammar (Melamed 2003), but also includes information of the target language in a consistent fashion to produce the best possible parse for the target language. The authors show that a joint model that uses a PCFG on the source side, small annotated parses on target side and a translation model for both the languages produces better and accurate parses when compared to a PCFG parser trained on a small amount of annotated parses alone. In particular, they factor a bilingual syntax model down to the product of two monolingual models. They further replace the original generative model with a discriminative model, with the underlying parsing algorithm unchanged. In their bilingual parser, the English and Korean parses are connected through word-to-word translational correspondence links or word alignment. The bilingual parser only deals with one-to-one mappings. The authors suggest using a union graph(Smith and Smith 2004) to relax the restriction and also reduce sparsity in the alignment. However, they also point out that this may be computationally expensive. Recently, Chiang and Rambow (2006) apply synchronous grammars based projection to Arabic dialects and Modern Standard Arabic (MSA), but they use explicit linguistic knowledge instead of a trained translation model that requires a parallel corpus.

4.3 Bayesian Grammar Models

A Bayesian grammar model provides a general method for obtaining parameters of transfer models without specifying transfer grammars. Jansche (2005) proposes a Bayesian projection model for transferring phrase structure trees. The basic goal is to infer target-language parse trees given source-language parse trees through a Bayesian statistical model (Figure 7). In this model, only the source-language parse trees are observed. Target language parse trees are treated as hidden variables. The model is decomposed into a target-language language model and a transfer model. The target-language language model is built from unannotated target-language text. It is used to infer target-language parse trees (T_i) from the target language side. The parameters of the target-language language model Λ are drawn from a Dirichlet distribution with hyper-parameters λ. The transfer model assigns probability to a source-language parse tree given the target-language parse tree. The parameters of the transfer model Ξ are drawn from another Dirichlet distribution with hyper-parameters ξ. Finally, the whole model specifies a joint probability over the source- and target-language parse trees and the model parameters. Hence, given a set of source-language parse trees, the probabilities of the target-language parse trees can be inferred from the model.

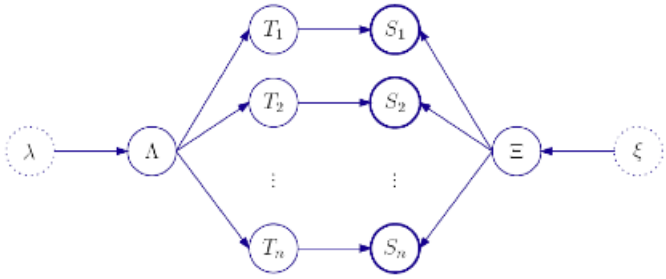

Figure 7
Treebank transfer model. Adapted from Jansche's slides on Treebank Transfer

Jansche's model provides a general technique for transferring annotations, which does not require alignment information and language-specific observations. Also, it gives an interesting explanation of syntax projection problems, where the annotations in target language are hidden variables which are expected to be recovered from the observations of source-language annotations.

5. Heuristic-Based Approaches to Syntax Projection

Heuristic-based approaches usually use some kind of parallel corpus with correspondences or alignments for transferring syntax. They also have an implicit or explicit notion of "direct correspondence assumptions" on the syntax under which the transfer is done. These approaches could broadly be summarized to consist of the following three phases:

Annotation	First identify the source units that are to be transferred. The source language text can be manually annotated or a tool can be used to annotate the text.
Transfer	The transfer of annotations takes place in this phase. The usage of some sort of correspondences between the words in the parallel sentence pairs is pre-identified. The quality of the correspondences decides the accuracy of the transfer. All transfers have some sort of "direct correspondence assumption" associated with them.
Postprocessing	Due to the syntactic divergences of the two languages, projection may produce noisy annotated data for the target language. Therefore in order to improve the quality of the data produced and to induce more robust tools from the data, a postprocessing phase is required. This phase incorporates and respects the target language syntactic constraints that may have been violated during transfer.

5.1 Projection via Word Correspondences

Most of the heuristic based approaches have roots from the work in word sense disambiguation (Resnik and Yarowsky 1999; Diab and Resnik 2001). But it was Hwa, Resnik, and Weinberg (2002) that introduced and then formalized (Hwa et al. 2005) the

Figure 8
Base noun phrase projection

assumption underlying in these models as the "Direct Correspondence Assumption" or DCA. The authors used it originally for dependency relation projection in (Hwa et al. 2005). Considering these approaches in retrospection, one can see that this assumption is quite valid for most of the heuristic based approaches to syntax projection. We borrow the term DCA and generalize the definition to any assumption used in syntax projection that is made for direct mapping. In individual lexical annotation projections, the common assumption is the annotations tend to be the same on two sides of the alignment. In flat sequential structures, one example of a direct correspondence assumption in noun-phrase bracketing is that a noun phrase in one language tends to remain an unbroken sequence when translated into another language (Yarowsky, Ngai, and Wicentowski 2001). Figure 8 shows an example of English noun phrases being projected to Chinese noun phrases. All the noun phrases in this example remain contiguous through projection. DCAs usually come from empirical studies of phenomena in bilingual corpora (Fox 2002). They are the basis and start for most of the heuristic based approaches to syntax projection. However, DCAs also tend to create very noisy annotations for target language because they are too simple and deterministic given the complexity of real languages. Thus, probability models are usually used on top of DCA for projection robustness (Yarowsky, Ngai, and Wicentowski 2001).

Unlike the grammar-based approaches discussed in previous Section 4, which are relatively new and being applied recently, the heuristic based approaches have been successfully applied to most of the syntax projection tasks. In this section we particularly discuss the work applied to POS tagging, noun-phrase bracketing, syntactic parsing and semantic role labeling, which raises interesting research challenges.

Yarowsky, Ngai, and Wicentowski (2001) discuss the experiments performed in inducing multilingual text analysis tools like POS taggers, base noun-phrase taggers, morphological analyzers, named-entity taggers and the like. The common underlying algorithm for all the tasks is to first word align the corpus using automatic probabilistic alignment algorithms and then reliably project syntax using the word alignment as a bridge. As already discussed, the two main hindrances to all these approaches are noisy word alignment due to lack of sufficient parallel data and syntactic divergences between the languages. Yarowsky, Ngai, and Wicentowski (2001) note that directly projecting the POS tags to a second language and training a tagger does not result in a very useful and accurate tagger. Therefore they discuss intelligent algorithms for training and inducing multilingual tools for separate annotation tasks. For a POS tagger, part of their strategy is to separate the tag sequence model $p(T)$ from the lexical model $p(W|T)$ and train each on varying amounts of data. The authors only choose data with higher alignment confidence. Cucerzan and Yarowsky (2002) further improve the robustness by incorporating contextual agreement to relax the strict Markovian assumption in POS

tagging. In particular, they check gender consensus in a relatively narrow window for Romanian, and the window size is chosen based on empirical studies of the gender-agreement ratio between a tagged word and other gender-marked words in context. Readers are encourage to read (Yarowsky, Ngai, and Wicentowski 2001) for details on other tasks, but in here we try to summarize the effort in the noun-phrase bracketing task.

The task of noun-phrase bracketing is to extract base noun-phrase structures from sentences. If we have aligned data, direct projections can be applied. The basic motivation for noun-phrase bracketers is that individual noun phrases tend to cohere sequentially. This means that a noun phrase in a language will remain an unbroken sequence when translated into another language, although the word order may vary. This assumption has also been supported elsewhere (Fox 2002; Koehn and Knight 2003). Yarowsky, Ngai, and Wicentowski (2001) also discuss the induction of an noun-phrase bracketing tool using the data obtained by syntax projection. The algorithm proceeds by first obtaining noun-phrase bracketed source-side data and then using the best word alignment for the parallel sentence pairs. The subscript of the noun phrase on the source-side is projected onto the target language sentence. The authors also observe that most of the noun phrases have a contiguous span on the target side and that any sort of interleavings in the target-side span of the noun phrase is only due to alignment errors. Figure 4 gives an example where this kind of a direct correspondence assumption fails. Therefore they also drop the data obtained from less confident word alignments to get better quality annotated data for training a standalone analyzer.

5.1.2 Dependency and Phrase Structure Trees. One of the difficult problems in natural language processing is syntactic parsing. Supervised methods for training parsers usually require an immense amount of annotated resources, which demands large human efforts. As such, it becomes difficult to build parsers for resource-poor languages. Hwa et al. (2005) discusses the feasibility of a "projection" based approach to create annotated resources for various languages and train statistical parsers on top of them. In particular, the paper explores and focuses on two important aspects: first, inferring complex structures like parse trees for a second language based on resource-rich monolingual data, parallel corpus and minimum human intervention; second, training high-quality parsers from noisy projections. The authors choose to work with dependency trees for the task of projection.

The authors also formalize the DCA that they make in order to deal with projection of complex tree structures. Given a pair of sentences E and F which are 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 $Tree_F$. In the example shown in Figure 9, the English word "got" is the parent of the word "gift". Also, "got" maps to the fifth Chinese word "mai", and "gift" maps to the eighth Chinese word "li-wu". So in the Chinese sentence, "mai" is the parent word of "li-wu". Under the assumption, the projection of the dependency trees is made using the word alignment as a bridge. For most languages, a post-projection transformation phase is required to deal with the monolingual idiosyncrasies of the language. For example, Chinese verbs are often followed by an aspectual marker that is not realized as a word in English. These require correction rules made by human inspection and analysis. The paper discusses experiments of creating parsers for Spanish and Chinese languages when projecting from English. The authors demonstrate that the initial DCA followed by post corrections enables them to seed and train parsers that yield about 67% F-scores for Chinese and

Figure 9
DCA in a dependency relation projection adapted from Hwa et.al (2005): Bootstrapping parses for resource poor languages

70% for Spanish in a constrained scenario and observe a drop of only 10% when working with large parallel corpora. F-score is an accuracy metric, which will be defined in more detail in section 6.

One of the major hindrances of projection for approaches like Hwa et al. (2005) and Yarowsky, Ngai, and Wicentowski (2001) are the low quality of word alignment. While Yarowsky, Ngai, and Wicentowski (2001) address this problem by redistributing the parameter values, Hwa et al. (2005) apply post-projection transformations to adjust the projections to improve the quality of the annotations. (Xi and Hwa 2005) in particular address the same problem in a slightly different way. Instead of completely projecting the data and deal with noisy data, the authors assume a small set of annotated data available for the resource-poor non-English language. This is similar in spirit to most bootstrapping algorithms that start with seeded data. The basic approach is to train two separate models from two different data sources. The first model is trained from a large corpus of automatically tagged data. The data is created by projection on the lines of Yarowsky and Ngai (2001). The second model is trained from a much smaller human-annotated corpus, where the set of sentences were automatically selected to improve the word coverage. Both the models are then combined into a single model via a back-off language model. The authors apply the approach to the POS tagging problem and discuss results that are better than either of the two approaches independently.

5.1.3 Semantic Role Labeling. (Padó and Lapata 2005) discuss an approach to projecting semantic role information across linguistic units on both sides of the language. Following the DCA paradigm, the projection takes place in three phases. Firstly, the source and target sentences are represented as sets of units U_s and U_t. These could be any linguistic constituents, usually phrase structure units. The semantic role assignment on the source-side is a function: R $\rightarrow$ (2^{U_s}) from roles to the set of source units. Next, constituent alignments are obtained between the two sides as another function: $U_s \times U_t$ $\rightarrow$ R . For robustness, only content words are used in the similarity calculation. Finally a decision procedure uses the similarity function to do the constituent mapping between the two sets of units. Once the mapping is completed, the role projection is just the transfer via constituent mapping links from the source to the target language. Two main contributions are the choice of linguistic units and the unit mapping algorithm. The linguistic units—usually phrase structure units—perform better than words as units.

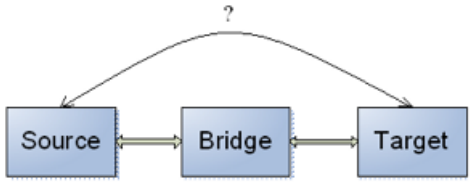

Figure 10
Bridge translation model

The authors also show the effectiveness of their constituent alignment algorithm, which performs about 0.65 F-score while matching phrase constituents.

Padó and Lapata (2006) further propose methods to solve the main challenge in Padó and Lapata (2005), which is finding the optimal mapping of the linguistic units on the source and target sides. The authors relax the independence assumption taken earlier that the alignment decision of two constituents is taken independently of the other constituents. They investigate well-understood global optimization models that suitably constrain the resulting alignments. Padó and Lapata (2006) model constituent alignment as a minimum-weight bipartite edge cover problem. Each of the set of units is a vertex set that is connected completely with all other units in the other set. The edge weights represent the dissimilarity between the vertex pairs. The problem now is to identify the minimum edge cover, which was solved using well-known algorithms. Besides matching constituents reliably, poor word alignments are a major stumbling block for accurate projections. Similar to other approaches addressed in this section, the authors also address this concern by proposing a novel filtering technique as a preprocessing stage. As part of the preprocessing to reduce the uncertainty of the tree, they remove extraneous constituents, like the non-content words or the words that remain unaligned. Also unlike Padó and Lapata (2005), the authors now use linguistic knowledge which states that not all words in a sentence are equally likely to be semantic roles. They give priority to children of the predicate and also constituents that do not have a sentence boundary between them and the predicate.

5.2 Projection using Bridge Languages

Bridge transitions are often used for filling gaps of alignments and thus guide new discoveries for missing relationships. Correspondence assumptions here are used in multiple pairs of languages, rather than two. We have seen in Section 3 that gaps in French morphology induction can be filled by English morphology links. In that example, English root-inflection relations serve as bridge links to morphology projection. Sometimes, a third language serves as a bridge to provide more clues for source-target syntax projection. This "third language" is also called the "bridge language".

Mann and Yarowsky (2001) propose methods for translation lexicons induction via bridge languages. The idea comes from the observation that words in translation lexicon pairs tend to have similar surface forms if they are from the same language family. Unlike other syntax projection methods, Mann and Yarowsky (2001) do not require aligned text. Rather, they only use a dictionary for mapping between the source language and bridge language. And the mapping from the bridge language to the target language is resolved by a probabilistic cognate model, where "cognate" refers to pairs of words that are similar both in meaning and surface form. For example, the lexicon annotation

projection from English to Portuguese is decomposed into two steps: first map English lexical entries to Spanish via English-Spanish dictionary, then map Spanish lexicons to Portuguese though probabilistic cognate models. Obviously the performance of the model depends on the similarity of the bridge language and the target language. The authors proved this intuition by experimental results. Given a bridge-target language pair, the performance of the cognate model depends on string distance measures. The authors compared three distance measures: edit distance (also called Levenshtein distance), a distance function learned from stochastic transducers, and a distance function learned from a hidden Markov model. Results show that weighted Levenshtein distance (weights are assigned to string-edit operations) gives the best accuracy.

One problem of the cognate model is that the assumption of equivalence on similarity of meaning and similarity of word surface form does not always hold. In other words, some correct mapping may have a lower similarity score than some false ones that happen to have a closer distance. In order to solve this problem, Schafer and Yarowsky (2002) propose seven complementary similarity models to capture true mappings and filter out the false ones. In addition to string similarity, these similarity models evaluate the similarity of context, time distribution, word frequency, and burstiness statistics. The final combination of the eight models gives an improved accuracy on English-Serbian test sets than the previous work done by Mann and Yarowsky (2001).

6. Evaluation

Most syntax projection models perform projection from one language to another in order to train and induce multilingual analysis tools (Yarowsky, Ngai, and Wicentowski 2001) for the target language. Some others perform a projection in order to build lexical resources in the target language (Diab and Resnik 2001). Therefore the evaluation of syntax projection depends on two main issues — the quality of the annotated data produced by projection and the quality of the tools that are induced. This leads to two different strategies for evaluation which we discuss in this section. Before that, we will first discuss another practical evaluation metric concerning resource prerequisites.

6.1 Data and Tools

One practical evaluation metric for a syntax projection model is the total human efforts and resources required for gathering the prerequisite data (Cucerzan and Yarowsky 2002). Most sequential and hierarchical annotation projection models require parallel texts and annotated data for source languages. These resources are especially important for heuristic based methods, where the alignment information is the basis of correspondence relations. Lexical annotation projections sometimes only need a bilingual dictionary as parallel data (Cucerzan and Yarowsky 2002). The required annotation and alignment can be human created. They can also be generated automatically from existing tools. For example, to obtain POS information for a source language, we can use a POS tagger. To obtain alignment information for parallel texts, we can use word alignment tool such as GIZA++ (Och and Ney 2000). Human knowledge for languages is also involved as a necessary resource for some models. For example, human-guided data filtering is a common technique used for preprocessing or postprocessing. In general, fewer prerequisites on resources and human efforts would be preferred when evaluating syntax projection models.

6.2 Strategies

6.2.1 Accuracy Metrics. When gold standard annotated data is available for the target language, one can compare the output produced by syntax projection with the gold standard for accuracy metrics. The definition of accuracy is different for different tasks. For the case of individual syntax and flat syntax structures like POS tagging and noun-phrase bracketing, the measures can be precision and recall. Precision and recall in syntax projection can be defined as below:

$$\text{Precision} = \frac{|\text{gold standard} \cap \text{total projections}|}{|\text{total projections}|}$$

$$\text{Recall} = \frac{|\text{gold standard} \cap \text{total projections}|}{|\text{gold standard}|}$$

$$\text{F-measure} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Precision + Recall}}$$

For example, Hwa et al. (2005) evaluate the accuracy of projection of treebank parses by comparing the precision and recall over human-annotated parse tree data. And Yarowsky, Ngai, and Wicentowski (2001) compare the accuracy of noun-phrase bracketing and POS tagging in a similar way.

6.2.2 Application-Focussed Evaluation. In application-focussed evaluation, syntax projection models are evaluated indirectly by specific tasks they are applied to or the effectiveness of the tools that are induced from the outcome. Evaluation of multilingual analysis tools is very often done by comparing their output on unseen test data using accuracy metrics mentioned above such as precision and recall. Sometimes the outcome of syntax projection is directly applied to downstream problems in natural language processing like machine translation (MT) (Quirk, Menezes, and Cherry 2005; Xia and McCord 2004) or word sense disambiguation (Diab and Resnik 2001). In such cases, the improvement in the specific task is evaluated as a quantifier of the syntax projection technique. Syntax-based approaches in statistical machine translation (SMT) are now making extensive use of the idea of syntax projection either to build syntax-driven translation models or learn translation rules from parallel corpus (Galley et al. 2004). For a detailed reading on syntax and MT, the readers are encouraged to read Ahmed and Hanneman (2006).

7. Applications of Syntax Projection

One direct application of syntax projection is to create annotated data for resource-poor languages, thus drive more active language research for these languages. This also enables us to apply existing structured model training techniques to induce multilingual tools. There is also recent interest in the area of improving word alignment by using syntactic annotations for one side of the corpus (Lin and Cherry 2003; DeNero and Klein 2007)(Lopez and Resnik 2005). All these methods reduce improper alignments by softly

enforcing the syntactic divergences that were trained by observing the corpus along with the syntactic information of one side of the parallel corpus.

Another application which is not explicit is that projection provides a tool for the linguists to understand a broad variety of languages. For example, the Language Navigation project at Carnegie Mellon University, looks at how feature structures and syntactic structures behave across various languages. The insights from syntax projection can also be directly applied to benefit the core problems like MT. In this regard, Mukerjee, Soni, and Raina (2006) perform a syntax-projection-focused experiment to study the complex predicates (CP) in Indian languages. CPs are very common in the Indo-Aryan language family. They are multi-word complexes functioning as a single verbal unit. This includes adjective-verb, noun-verb, adverb-verb and verb-verb composites. Since most of the Indo-Aryan languages are resource-poor, we need the help of projecting POS from English. The method requires parallel corpus of English and Hindi.

Ideas of bridge language based projection techniques discussed in section 5.2 have also been used in statistical SMT (Koehn, Och, and Marcu 2003)(Brown et al. 1993). In state-of-the-art SMT models, high quality phrase tables are essential for a language pair for better quality translation. For a vast majority of language pairs, we do not have sufficient data to train SMT models. Projection models use bridge languages to create phrase tables where parallel corpus does not exist, thus enabling us to build machine translation systems for more language pairs. Such an approach is successfully demonstrated in Utiyama and Isahara (2007). Even though there are large volumes of parallel data for Chinese-English and Arabic-English, there are few resources for Chinese-Arabic pair. Observing this, the authors propose a method using a pivot language such as English to bridge the source and target languages. For the Chinese-English-Arabic example, we assume that we have a Chinese-English phrase table and an English-Arabic phrase table, based on which we can construct a Chinese-Arabic phrase table. Phrase translation probabilities and lexical translation probabilities for the Chinese-Arabic pair need to be estimated by the assistance of English-X translation model, where X stands for a target language such as Chinese or Arabic. For sentence translation, two independently trained SMT systems (Chinese to English and English to Arabic) are used. The idea is to first translate a Chinese sentence into several English sentences, and then translate those English sentences with highest score into Arabic. There are other applications for syntax projection. For example, syntax projection is also known to automatically induce information extraction systems, where the information extraction system is trained from annotated data obtained by syntax projection (Riloff, Schafer, and Yarowsky 2002). We will not enumerate all the applications for syntax projection, but it should be clear to the reader that syntax projection in general is a useful technique for multilingual learning and many other applications can benefit from it especially in the resource-poor language scenario.

8. Conclusion

We have seen a swell of interest in multilingual syntax learning over the past decade. One major goal of multilingual syntax learning is to learn monolingual syntax with the help of other languages. This help mainly comes from three different kinds of resources. First, a resource-poor language can obtain annotations from a resource-rich language through syntax projection. For example, we can generate dependency trees through projection (Hwa et al. 2005). Second, a bridge language can be used for filling gaps of a resource-poor language and a resource-rich language. For example, we can use Spanish to help projecting annotations from English to Portuguese (Mann and Yarowsky 2001).

Third, a resource-rich language can also benefit from syntax projection. For example, we can disambiguate English words by checking with the translation lexicon in another language (Resnik 2004). In our survey we summarized the efforts developed to project syntax from a source language to a target language. Although we categorize current projection techniques into grammar-based methods and heuristic-based methods in this survey, these two methods are closely related to each other. Grammar-based methods are also heuristic-based, because the transition rules also specify a correspondence assumption between the syntax structures of the two languages. On the other hand, heuristic-based methods are also grammar-based, because the correspondence assumptions can be treated as a transition grammar, and the postprocessing procedure serves as a set of grammar rules for the target language. This is more obvious in the example of Hwa et al. (2005), where the dependency tree projection model is decomposed into a transition model based on correspondence rules and a postprocessing phase based on target-language-specific filtering rules. The main difference that separates these two paradigms is that grammar-based models have a clearer and more general formalism for the rules that include the source and target side linguistic knowledge together, while heuristic-based models concentrate on transferring the annotations from the source side and postprocessing to conform with the target language constraints. For projection algorithms, deterministic methods combined with probability models are shown to provide robust performance in syntax projection problems. For example, we have presented that in (Schafer and Yarowsky 2002) for inducing translation lexicons, Levenshtein distance similarity measure (a deterministic measure) is combined with probability measures, such as time distribution, word frequency, and burstiness statistics, to generate robust output.

We provided an overview of multilingual syntax projection problems. We introduced major challenges for syntax projections on different syntactic structures. We gave a discussion of two main techniques used for syntax projections (grammar-based methods and heuristic-based methods). And we presented the evaluation metrics, and finally application for syntax projection. Through the survey, we avoid technical details for the methods. For these details, the readers are encouraged to read related papers that we refer to in our report. We focussed on the main ideas behind various syntax projection methods, and we hope to have given a comprehensive view of this interesting research area.

9. Acknowledgements

We are thankful to Greg Hanneman for proofreading the report and providing valuable suggestions.

References

Ahmed, Amr and Greg Hanneman. 2006. Syntax based statistical machine translation: A review. http://www.cs.cmu.edu/ nasmith/LS2.F06/.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chiang, David and Owen Rambow. 2006. The hidden TAG model: Synchronous grammars for parsing resource-poor languages. In *Proceedings of the Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms*, pages 1–8, Sydney, Australia, July. Association for Computational Linguistics.

Cucerzan, Silviu and David Yarowsky. 2002. Bootstrapping a multilingual part-of-speech tagger in one person-day. In *COLING-02: proceeding of the 6th Conference on Natural Language Learning*,

pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.
DeNero, John and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.
Diab, Mona and Philip Resnik. 2001. An unsupervised method for word sense tagging using parallel corpora. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 255–262, Morristown, NJ, USA. Association for Computational Linguistics.
Fox, Heidi J. 2002. Phrasal cohesion and statistical machine translation. In *EMNLP '02: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, pages 304–3111, Morristown, NJ, USA. Association for Computational Linguistics.
Galley, Michel, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In Susan Dumais; Daniel Marcu and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 273–280, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.
Hwa, Rebecca, Philip Resnik, and Amy Weinberg. 2002. Breaking the resource bottleneck for multilingual parsing. In *Proceedings of the Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*.
Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325, September.
Jansche, Martin. 2005. Treebank transfer. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 74–82, Vancouver, British Columbia, October. Association for Computational Linguistics.
Klementiev, Alexandre and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 817–824, Morristown, NJ, USA. Association for Computational Linguistics.
Koehn, Philipp and Kevin Knight. 2003. Feature-rich statistical translation of noun phrases. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.
Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edomonton, Canada, May 27-June 1.
Lin, Dekang and Colin Cherry. 2003. Word alignment with cohesion constraint. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 49–51, Morristown, NJ, USA. Association for Computational Linguistics.
Lopez, Adam and Philip Resnik. 2005. Improved hmm alignment models for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 83–86, Ann Arbor, Michigan, June. Association for Computational Linguistics.
Mann, Gideon S. and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.
Melamed, I. Dan. 2003. Multitext grammars and synchronous parsers. In Marti Hearst and Mari Ostendorf, editors, *HLT-NAACL 2003: Main Proceedings*, pages 158–165, Edmonton, Alberta, Canada, May 27 - June 1. Association for Computational Linguistics.
Mukerjee, Amitabha, Ankit Soni, and Achla M Raina. 2006. Detecting complex predicates in hindi using pos projection across parallel corpora. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 28–35, Sydney, Australia, July. Association for Computational Linguistics.
Och, Franz Josef and Herman Ney. 2000. A comparison of alignment models for statistical machine translation. pages 1086–1090, Saarbrucken, Germany, August.
Padó, Sebastian and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 859–866, Morristown, NJ, USA. Association for Computational Linguistics.

Padó, Sebastian and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 1161–1168, Morristown, NJ, USA. Association for Computational Linguistics.

Probst, Katharina. 2003. Using 'smart' bilingual projection to feature-tag a monolingual dictionary. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 103–110, Morristown, NJ, USA. Association for Computational Linguistics.

Quirk, Chris, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 271–279, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Resnik, Philip. 2004. Exploiting hidden meanings: Using bilingual text for monolingual annotation. In *CICLing*, pages 283–299.

Resnik, Philip and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistiscs*, 29(3):349–380.

Resnik, Philip and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Riloff, Ellen, Charles Schafer, and David Yarowsky. 2002. Inducing information extraction systems for new languages via cross-language projection. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Ryan T. McDonald, Fernando Pereira, Kiril Ribarov and Jan Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *HLT/EMNLP*.

Schafer, Charles and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *COLING-02: proceeding of the 6th conference on Natural language learning*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Shieber, Stuart M. and Yves Schabes. 1990. Synchronous tree-adjoining grammars. In *Proceedings of the 13th Conference on Computational Linguistics*, pages 253–258, Morristown, NJ, USA. Association for Computational Linguistics.

Smith, David A. and Noah A. Smith. 2004. Bilingual parsing with factored estimation: Using english to parse korean. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 49–54.

Utiyama, Masao and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York, April. Association for Computational Linguistics.

Wu, Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Xi, Chenhai and Rebecca Hwa. 2005. A backoff model for bootstrapping resources for non-english languages. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 851–858, Morristown, NJ, USA. Association for Computational Linguistics.

Xia, Fei and Michael McCord. 2004. Improving a statistical machine translation system with automatically learned rewrite patterns. In *COLING '04: Proceedings of the 20th International Conference on Computational Linguistics*, page 508, Morristown, NJ, USA. Association for Computational Linguistics.

Yarowsky, David and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Yarowsky, David, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT '01: Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.