Toward Machine Translation with Statistics and Syntax and Semantics

Dekai Wu

Hong Kong University of Science & Technology
Human Language Technology Center
HKUST, Department of Computer Science & Engineering, Hong Kong
dekai@cs.ust.hk

***Abstract*—In this paper, we survey some central issues in the historical, current, and future landscape of statistical machine translation (SMT) research, taking as a starting point an extended three-dimensional MT model space. We posit a socio-geographical conceptual disparity hypothesis, that aims to explain why language pairs like Chinese-English have presented MT with so much more difficulty than others. The evolution from simple token-based to segment-based to tree-based syntactic SMT is sketched. For tree-based SMT, we consider language bias rationales for selecting the degree of compositional power within the hierarchy of expressiveness for transduction grammars (or synchronous grammars). This leads us to inversion transductions and the ITG model prevalent in current state-of-the-art SMT, along with the underlying ITG hypothesis, which posits a language universal. Against this backdrop, we enumerate a set of key open questions for syntactic SMT. We then consider the more recent area of semantic SMT. We list principles for successful application of sense disambiguation models to semantic SMT, and describe early directions in the use of semantic role labeling for semantic SMT.**

I. Introduction

Ever since the successes of the neo-statistical machine translation movement began to reorientate the field back toward quantitative learning approaches some twenty years ago, tensions—real and imagined—have persisted between methodologies that prioritize statistical modeling versus structural, symbolic modeling.

Statistical machine translation, or SMT, resurrected the classic ideas of Weaver (1949) in positing a noisy channel process for modeling translation, while exploiting a half-century of vast advances in computing hardware. To a fresh generation of NLP researchers, SMT offered an irresistable trio of attractions long enjoyed by the speech recognition and pattern recognition communities, and painfully absent in the NLP methodology of the time: (a) an empirical research process grounded in the scientific method, (b) numerically weighted decision models well suited for integrating indirect partial evidence from multiple disparate clues, often grounded in Bayesian foundations, and (c) machine learning techniques for breaking the knowledge acquisition bottleneck, often grounded in information theory.

Yet even as we "neo-stats" have come to dominate machine translation research in the intervening years, oversimplistic representational assumptions continue to dog most SMT models. This yields mistranslations that, while superficially fluent, are often jarringly inadequate. The most cursory error analysis instantly reveals a preponderance of obvious syntactic and semantic errors. Language pairs that are very different, like Chinese and English, are particularly sensitive to such problems.

It has proven intriguingly difficult to avoid throwing the baby out with the bath water. The careful training and optimization of statistical MT models makes them remarkably counterbalanced: attempts at incorporating syntactic or semantic models so as to improve translation adequacy tend to meet, more often than not, with degraded fluency. At the same time, MT evaluation metrics such as BLEU (Papineni *et al.*, 2002) that have been dominant in recent years reward fluency at least as highly as adequacy. This creates little practical incentive to invest time attacking the underlying problem compared with, say, fine-tuning model parameters, or engineering ever larger systems capable of memorizing still more hundreds of millions of phrase translations.

Nevertheless, many perceived differences between statistical and symbolic translation modeling are illusory. A history of MT paradigms in Wu (2005) factors out some of the artificial distinctions by plotting various approaches in a three-dimensional space of possible machine translation models, shown in Figure 1. Irrespective of the extent to which a model employs statistics (the first dimension), several design choices must be made.

One design choice (the second dimension) is when the inductive steps—generalization, adaptation, learning—are performed. While this might appear to be a purely implementational choice if we were to disregard issues of computation time and space, the reality is that computation is always resource-bounded, so this design choice actually produces different models. In MT models that are more **example-based**, induction is largely done at runtime during testing, by adapting fragments of memorized sentence translations during translation decoding. In MT models that are more **schema-based**, induction is mainly performed in advance, by capturing abstract generalization patterns via either automatic training or manual construction.

Another design choice (the third dimension) is the degree to which the representational foundation allows for **compositional** structures or rules, versus being restricted to solely manipulating "flat" chunks in the form of **lexical** strings or

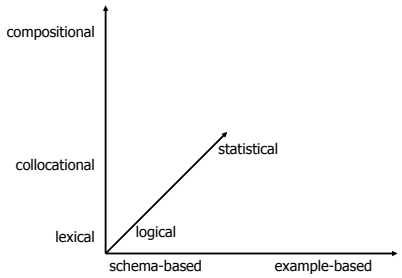


Fig. 1. Three independent dimensions of MT model space.

segments. The principle of compositionality lies at the heart of **syntactic** modeling. Yet even today, much of SMT still employs flat lexical modeling—or only the degenerate case of compositionality in its purely flat form, where lexical strings hold segments variously referred to as **compound words**, **phrases**, **collocations**, or **multiword expressions**—largely out of concern for the computational complexity of translation decoding and training. A key question, then, is how much compositional power to allow, in order to adequately capture the necessary generalization patterns, but without admitting excessive computational complexity.

To these design choices we will add one more: the extent of **semantic** modeling. The most glaring errors made by SMT systems arise from current inadequacies in modeling of the **context** within which lexical segments are translated. Of particular interest are recent lexical semantics models for word sense disambiguation and semantic role labeling, which attack problems that are largely orthogonal to those addressed by current SMT architectures. While work on syntax for SMT has had since Wu (1995a) to develop and now is extremely active—see for example Wu and Chiang (2007), Chiang and Wu (2008), and Wu and Chiang (2009)—serious work on semantics for SMT began much more recently (Carpuat and Wu, 2005) and is still at a much earlier stage.

Our consideration of these issues is set against the background of one of the more frustrating puzzles of the day: why are some language pairs—like Chinese and English—so much more resistant to machine translation than others? The US Defense Language Institute classifies both Arabic and Chinese in Group IV, the most difficult languages for English speakers to learn, and both Semitic and Sino-Tibetan languages evolved from separate origins as Indo-European. Yet the much more recent efforts on Arabic-English very quickly leapfrogged SMT accuracy for Chinese-English. Wu (2008) posits a **socio-geographical conceptual disparity hypothesis**: the Arabic and European worlds shared vast borders with a long history of *direct* trade and rule, unlike the Far East and Europe. Far more borrowing of concepts has occurred over the ages between Arabic and European languages, than between Chinese and European languages. Thus, there is a much higher chance that an idea expressed by an Arabic phrase can easily be translated into English via direct word/phrase substitution, because each component concept translates easily.

Regardless, our discussion below largely applies to all languages. But concrete Chinese-English examples illuminate the issues strikingly well, due to their vast differences. Chinese-English MT is an acid test; breaking through the current plateau in Chinese-English MT quality will demand that SMT incorporates much richer representational models to bridge the gap.

II. Syntactic SMT: What's in a word?

The past and future evolution of syntactic SMT approaches is most clearly discerned by contrasting **token-based**, **segment-based**, and **tree-based** models, respectively corresponding to the design choice between lexical, collocational, and compositional representations.

A. Token-based models

The simplest **lexical** SMT models, notably the IBM models of Brown *et al.* (1988), manipulate single lexical tokens using flat, non-compositional permutation models. These models are often referred to as being "word-based"—a rather ill-defined concept, given the murky status of "words" in languages written with no whitespace delimiters, like Chinese, where nearly every character can be meaningfully used by itself. The unfortunate nomenclature arose as a historical artifact of an early focus on Western European languages, particularly English and French.

The key distinction of this class of models is that the algorithms and translation lexicons make no attempt to translate via lexicon lookup for any segments longer than one single token.

The limitations of flat single-token approaches quickly become painfully apparent when applied to languages like Chinese, where no obvious whitespace boundaries can be used to delimit "words". Consider, for example, the following sentence pair:

Chi	管理局将会向财政司负责。
Gloss	*authority will to financial secretary hold responsibility.*
Eng	The authority will be accountable to the financial secretary.

Two approaches toward dealing with the Chinese string's lack of whitespace are possible—**token-based** versus **segment-based** models—and they often produce confusing conflicts of perspective and terminology.

The **token-based** approach assumes some heuristic preprocessor will chop the input Chinese string of characters into a sequence of multi-character "word" tokens such as 负责 (fùzé, roughly *accountable*). Under this approach, the heuristic preprocessing is typically called "**word tokenization**", reflecting the assumption that the resulting chunks are atomic tokens that will later be translated by lexicon lookup. The task of Chinese-English translation is then shoehorned into the same token-based SMT models developed earlier for European language pairs.

This strategy forces premature tokenization decisions that are often arbitrary, with consequences for translation accuracy: incorrect tokenization decisions lead to incorrect or suboptimal lexical translations. For example, tokenizing 负责 into a single token prevents it from being better translated in certain

contexts as two separate tokens, 负 (fù, roughly *hold* or *take* in this context) and 责 (zé, roughly *responsibility*), yielding *hold responsibility* or *take responsibility*.

B. Segment-based models

The alternative **segment-based** approach is less restrictive. Instead of prematurely forcing tokenization decisions, we simply recognize all individual Chinese characters as tokens. After all, individual Chinese characters can be used and translated meaningfully—just like "words" in European languages. When working with SMT for European languages, "words" are the atomic units that it would be unreasonable for SMT to hypothesize breaking into smaller pieces to be translated individually. Since it is almost always reasonable to hypothesize translating single Chinese characters, they are the closest analog to the atomic units considered "words" in European languages.

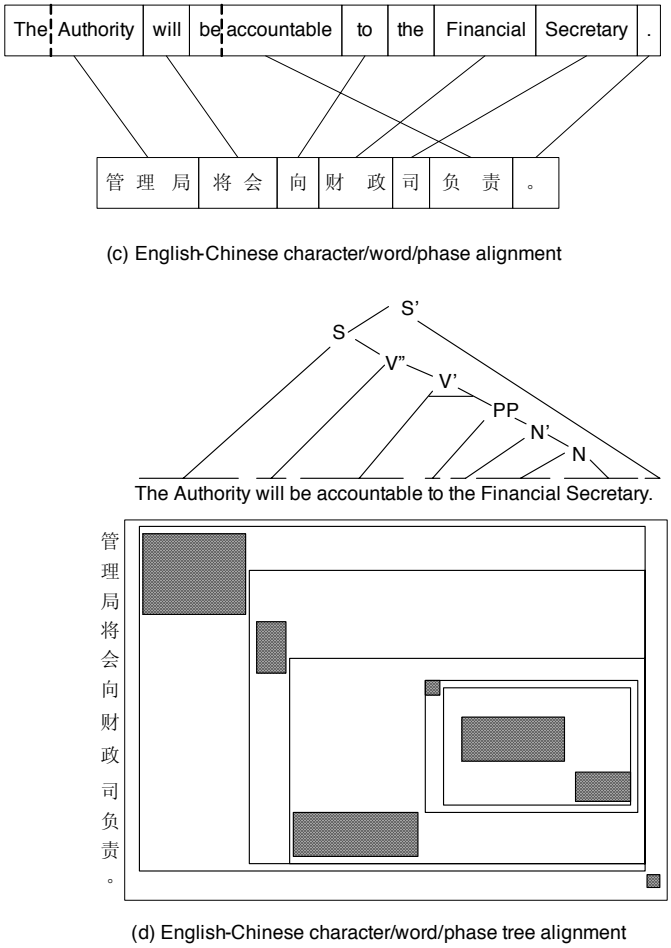
We then employ a **collocational** or *phrasal* translation lexicon, in which multi-token segments can be translated as a compound unit. Thus the lexicon simultaneously includes 负/*hold*, 责/*responsibility*, and 负责/*accountable*—just like in traditional translation dictionaries.

Under this approach, the decision as to what granularity of strings to chop the input Chinese string of characters into is typically called "**word segmentation**" rather than "word tokenization". This reflects the perspective that the "words" being translated by lexicon lookup are segments often containing multiple tokens. (The term "segment" implies multiple tokens—compare the use of terms like "sentence segmentation", "paragraph segmentation", or "clause segmentation", which are not referred to as "tokenization" since sentences, paragraphs, and clauses generally consist of multiple tokens.)

The key principle that distinguishes segment-based models is that they defer segmentation decisions until bilingual optimization decisions can be made at translation alignment or decoding time, instead of heuristically committing in advance to rigid boundaries during some monolingual preprocessing phase. This is called **translation-driven segmentation** (Wu, 1997) and is the core assumption underlying all phrasal translation models, including most **example-based MT** (EBMT) models (Nagao, 1984) and all **phrase-based SMT** (PBSMT) models (Och *et al.*, 1999).

Translation-driven segmentation is a key characteristic of the tree-based SMT models where it first developed—particularly the **inversion transduction grammar** (ITG) models as in Wu (1997) and Wu and Wong (1998), of which recent incarnations include Chiang (2007), Cherry and Lin (2007), Xiong *et al.* (2006), Xiong *et al.* (2009), or Haghighi *et al.* (2009) for example.

The success of translation-driven segmentation approaches over earlier single token-based models has not been limited to translation of Chinese. Even for European languages, lexicographers have long since given up any pretense that "words" are merely those strings delimited by whitespace. A huge proportion of entries in any English lexicon (or translation lexicon) are composite "words" variously referred to as **compound words**, **phrases**, **collocations**, or **multiword expressions**—flip to any random page to see entries like *each other*, *eager beaver*, *Eagle Scout*, *eared seal*, *earth science*, *earth station*, *east by north*, *Easter egg*, *Eastern Standard Time*, etc.—few of which can be reliably translated solely by their individual space-delimited tokens.



(c) English-Chinese character/word/phase alignment

(d) English-Chinese character/word/phase tree alignment

Fig. 2.

Machine translation models go further. Even segments that a lexicographer would not recognize are still often needed in the phrasal translation lexicon. For instance, our earlier example 负责 can also translate to *two* English words, *be accountable*, as in the segment-based alignment depicted in Figure 2(c). Many models do not even require the phrasal segments to correspond to complete linguistic constituents, relying instead on a language model to improve the grammaticality of the output translation.

Despite the improvement in representational power over token-based models, a phrasal segment consisting of multiple tokens still represents only a single, non-recursive level of composition—too limited to effectively model most syntactic patterns and constraints. Ungrammatical outputs are a frequent product of this class of models.

C. Tree-based models

To be able to model true compositional structure and properly capture long-distance dependencies, we need nested, recursive levels of composition. This was introduced to SMT in the stochastic **inversion transduction grammar** (ITG) models of Wu (1995a) and Wu (1997), a recent instantiation of which is the hierarchical phrase-based translation model of Chiang (2007).

The key distinction of most state-of-the-art tree-based models is that a single **transduction grammar** simultaneously models two languages. A few concepts are useful to define.

A **transduction** is a set of sentence translation pairs or **bisentences**—just as a language is a set of sentences. The set defines a relation between the input and output languages.

In the *generative* view, a transduction grammar generates a transduction, i.e., a set of bisentences—just as an ordinary (monolingual) language grammar generates a language, i.e., a set of sentences. In the *recognition* view, alternatively, a transduction grammar **biparses** or accepts all sentence pairs of a transduction—just as a language grammar parses or accepts all sentences of a language. And in the transduction view, a transduction grammar **transduces** (translates) input sentences to output sentences.

With such models, alignment becomes part of the biparsing process. Given a sentence pair, biparsing it produces the alignment dictated by the full biparse tree.

Larger segments are translated via **composition** of the translations of smaller segments. This composition process is recursive and stochastic. Figure 2(d) shows how our earlier sentence pair example can be biparsed (or transduced or generated) by a single **biparse tree**. At each node of the parse tree, choices are made about how the child nodes are to be permuted for the translation. For instance, the horizontal bar at the V' node is a shorthand indicating that the Chinese segment 负责 generated by the left child *be accountable*/负责 should be **inverted** with the Chinese segment 向财政司 generated by the PP right child (corresponding to *to the financial secretary*). The same nesting and permutation information can also be visualized with the matrix directly under the parse tree.

Whenever permutations come into the picture, exponential time and space complexities cannot be far behind. With compositional tree-based models, we now have the stochastic transduction grammar machinery to properly model long-distance dependencies. But if they come with infeasible computational complexities, we are no better off than before. This was one of the main factors that slowed widespread adoption of syntactic modeling into SMT. How can we gain sufficient compositional modeling power without excessive computational complexity?

III. Syntactic SMT: How much compositional power?

To address the question of how much compositional power to aim for in a tree-based model, it is useful to return to one of the fundamental principles of machine learning—the critical role of the **inductive bias** inherent in the assumptions of any learning model (Mitchell, 1997). The inductive bias consists of all *a priori* assumptions outside the training data. Without an inductive bias, no learning can be rationally justified. There are two main sources of inductive bias. A **search bias** is a preference for certain hypotheses over others; this follows from the *a priori* definition of an algorithm's **search strategy** and **objective criteria**, and does not place any hard restriction on what hypotheses can be enumerated in the course of the search. On the other hand, a **language bias** is a categorical restriction on the set of hypotheses that can be considered; this follows from the *a priori* definition of the **search space**. All learning systems intrinsically possess both search biases and language biases.

How, then, should we determine which inductive biases to embody within a model's language biases, as opposed to its search biases? All other things being equal, it is more efficient to formulate an inductive bias as a language bias, rather than a search bias. If we are fairly certain that some class of hypotheses will never be correct or optimal, then implementing a search bias that searches and then rejects those hypotheses clearly cannot be more efficient than forming an *a priori* language bias that eliminates those hypotheses from the search space in the first place.

However, the most widespread token-based models (IBM models) and segment-based models (phrase-based and example-based MT models) focus almost entirely on search biases. Their underlying generative models allow *all* permutations of the lexical translation units (tokens or segments), which is not much of a language bias. What few language biases they do employ—primarily finite window sizes for reordering, such as the "IBM constraints"—tend to be either relatively weak, or excessively harsh on long-distance dependencies. This places an extremely demanding burden on the heuristics implementing the search bias—resulting either in very inefficient search that wastes too much time in the wrong hypothesis regions, and/or very inaccurate search that fails to find correct or optimal hypotheses in the allotted time.

One major advantage of formal transduction models is that they offer a mechanism to impose strong language biases. But allowing biparse trees of nodes that generate arbitrary permutations of many children is a rather weak and ineffective language bias. It turns out there is a **hierarchy of equivalence classes for transductions**—just as there is Chomsky's hierarchy of equivalence classes for languages. Just as in the monolingual case, there is a tradeoff between generative capacity and computational complexity. Figure 3 summarizes both hierarchies, where the more expressive classes of transductions are orders of magnitude more expensive to biparse and train. The bilingual hierarchy is anchored on both ends by two familiar classes of transductions in widespread use for decades in many areas of computer science and linguistics.

At the upper end, we have the well-known equivalence class of **syntax-directed transductions**. Syntactic SMT systems that are entirely based on a pure unrestricted **syntax-directed transduction grammar** (SDTG) model (Lewis and Stearns (1968), Aho and Ullman (1969), Aho and Ullman (1972))—also recently called **synchronous CFG**—tend to

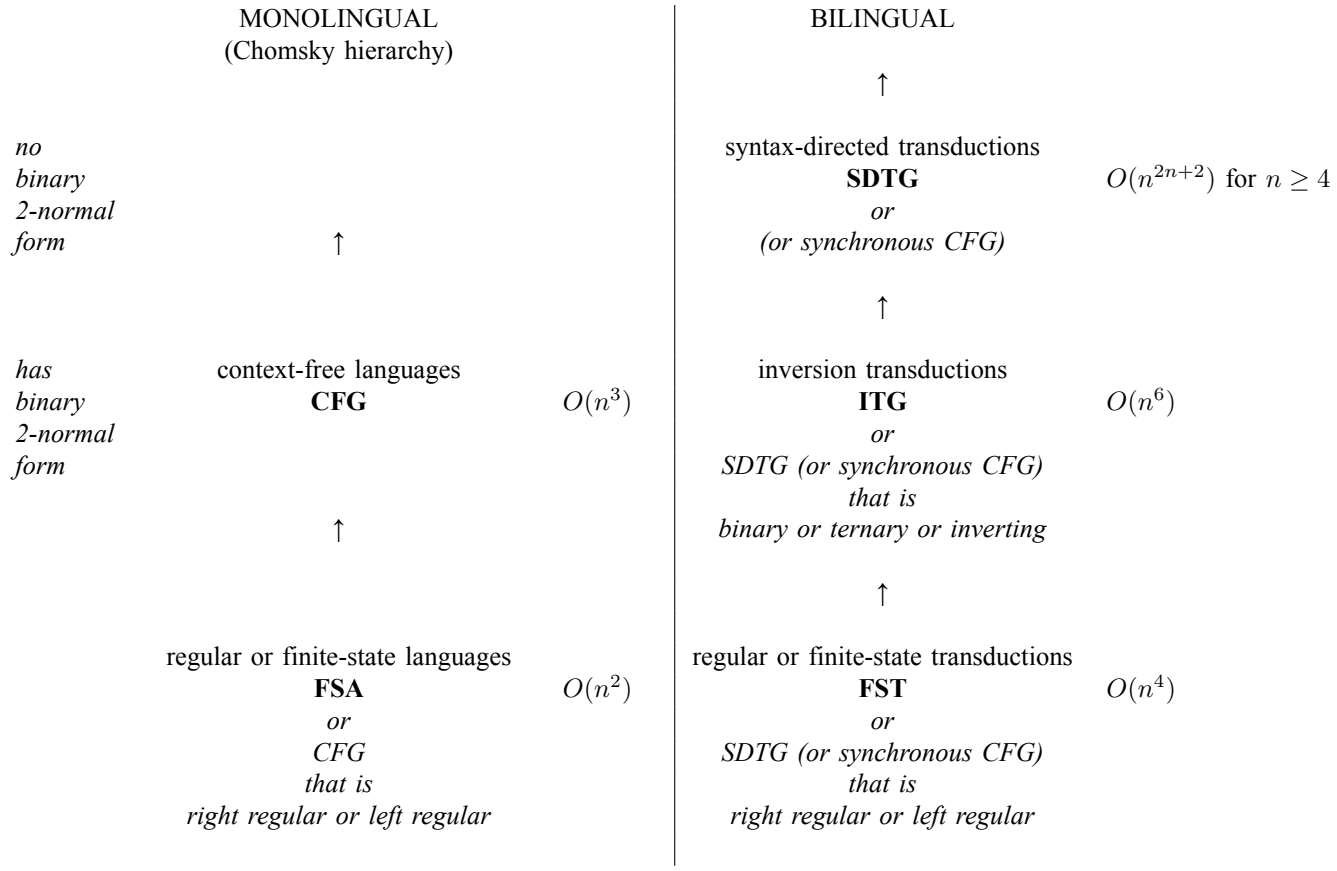


Fig. 3. Summary comparison of computational complexity for Viterbi and chart (bi)parsing, and EM training algorithms for both monolingual and bilingual hierarchies.

suffer from the weak language bias problem, resulting in inefficient and/or inaccurate search in training, decoding, or both. (Traditional rule-based MT models, like programming language compilers, usually attempt to circumvent such problems by carefully crafting their SDTGs by hand. Of course, this also leads to scaling difficulties.)

At the lower end, we have the equivalence class of **finite-state transductions**, which are the set of bisentences generated by some **finite-state transducer** (FST). It is possible to use SDTGs (or synchronous CFGs) to describe finite-state transductions by restricting them alternatively to the special cases of either "right regular SDTGs" or "left regular SDTGs". However, such characterizations overlook the key point—the equivalence class of finite-state transductions imposes a strong language bias, making the grammars orders of magnitude cheaper to biparse, train, and induce than with syntax-directed transductions—and also more accurate to induce for appropriate classes of problems.

In between lies the intermediate equivalence class of **inversion transductions**, which in recent years has seen wide use in nearly all state-of-the-art MT systems. The generative capacity and computational complexity for the **inversion transduction grammar** (ITG) class falls in between that of finite-state and syntax-directed transduction grammars; see Wu (1997) for detailed analysis of the expressiveness properties of ITGs, which are surprisingly flexible despite their strong constraints. The following ITG generates the tree from Figure 2(d), where square brackets are a shorthand for the straight permutation (0, 1) and angle brackets for the inverted permutation (1, 0):

S' → [S ./。]
S → [*The authority*/负责 V"]
V" → [*will*/将会 V']
V' → ⟨ *be accountable*/负责 PP ⟩
PP → [*to*/向 N']
N' → [*the*/ε N]
N → [*financial*/财政 *secretary*/司]

As with finite-state transductions, it is possible to use SDTGs (or synchronous CFGs) to describe inversion transductions by restricting them alternatively to the special cases of "binary SDTGs", "ternary SDTGs", or "SDTGs whose transduction rules are restricted to straight and inverted permutations only".

This means that ITGs are in a sense the closest bilingual analog of monolingual CFGs. All inversion transductions can be written in a binary **2-normal form**, like the example grammar above, just as all monolingual context-free languages can be written in a binary 2-normal form, such as Chomsky normal form. In contrast, SDTGs (or synchronous CFGs) do not admit binary 2-normal forms.

It also means that any SDTG (or synchronous CFG) of binary or ternary rank—i.e., that has at most two or three nonterminals on the right-hand-side of any rule—is an ITG.[1]

For example, the numerous systems that employ "binarized synchronous/ transduction grammars" reduce to the class of ITGs. One way this is often accomplished in practice is by applying the binarization algorithm of Zhang *et al.* (2006). This "down-converts" a less tractable SDTG (or synchronous CFG) by approximating it with an ITG that discards any transduction patterns that violate ITG constraints.[2] Translation speed and accuracy *improve* significantly as a consequence of this expressiveness restriction—a strong indication of the good fit of the ITG language bias to the domain of human language translation. Similarly, any grammar induced following the hierarchical phrase-based translation method (Chiang, 2007), which always yields a binary rank transduction grammar, is an ITG.

A key characteristic responsible for the success of such models is the strong language bias of inversion transductions, which makes the grammars orders of magnitude cheaper to biparse, train, and induce than with SDTGs (or synchronous CFGs).

This is encapsulated by the **ITG hypothesis** which posits a strong language universal constraint: *sentence translation between any two natural languages can be accomplished within the permutations allowed by ITG expressiveness*.[3] The conjecture that human sentence translations fall within the space of inversion transductions is analogous to the monolingual hypothesis that sentences of human languages are context-free—while certain limited classes of constructions can be found that in principle violate the constraints, these tend to be restricted cases, statistically rare, and easily handled via simple preprocessing. An overwhelming proportion of languages/ transductions are most effectively covered within efficient CFG and ITG constraints.

Over the years, numerous empirical results have borne out the ITG hypothesis to a surprisingly large extent, indicating significantly better fit to modeling translation between many human language pairs—with more efficient and/ or effective search for alignment as well as translation decoding, across English, French, Spanish, German, Swedish, Chinese, Japanese, Arabic, and numerous other languages. A few examples of these include Zens and Ney (2003), Zens *et al.* (2004), Gildea (2004), Wu *et al.* (2006), and Saers and Wu (2009).

IV. Syntactic SMT: Open questions

The success of tree-based models crystallizes a number of key open questions. While initial work toward each question has begun, much remains to be answered.

- *Language bias of ITGs or SDTGs (synchronous CFGs)?* The bulk of the evidence currently suggests that the space of inversion transductions is a very good fit to translation between human languages, and minimizes computational complexity while improving accuracy of alignment, translation, and grammar induction (which is why models that start with SDTGs (or synchronous CFGs) often apply the ITG binarization method of Zhang *et al.* (2006) to "down-convert" them). But is there an empirically demonstrable, statistically significant counterargument to the ITG hypothesis?
- *Generic or specific transduction grammars?* At the most generic extreme, simple **bracketing ITG** (BITG or BTG) models of Wu (1995a), which employ only two syntactic rules $X \rightarrow [XX]$ and $X \rightarrow \langle XX \rangle$ with a single undifferentiated nonterminal, are used solely for the language bias from their restrictions on permutations—often referred to as **ITG constraints**—and have proven very useful in many models. But other models employ and/or induce far larger, more specific transduction grammars, as shown in Figure 4. For example, detailed linguistic grammars are used to construct linguistic ITGs in Wu and Wong (1998).
- *Unsupervised grammar induction or linguistically constructed grammars?* Closely related to the previous question, unsupervised approaches often begin with BITG models since they require no *a priori* linguistic knowledge, whereas linguistically constructed grammars typically hold a fair amount of specific syntactic knowledge. Unsupervised approaches treat trees as entirely hidden structure; many such approaches do not care if the tree shapes end up being linguistically conventional. But many other blends are possible.
- *How are model parameters trained?* Reasonably efficient EM training methods are available for training large numbers of parameters in token-based IBM models (Brown *et al.*, 1993) as well as tree-based ITGs (Wu, 1995b). Maximum-entropy and minimum error-rate training methods are also available for tuning relatively small numbers of parameters in loglinear models used in many systems including segment-based PBSMT models (Och, 2003), but their stability is not very reliable. Can better methods for parameter estimation be developed?
- *Lexicalization?* Are the benefits ultimately worth the cost of fully lexicalizing ITGs—for example, as in the lexically parameterized ITGs of Huang *et al.* (2005), Zhang and Gildea (2005), or Zhang and Gildea (2006), the maximum-entropy BITG models of Xiong *et al.* (2006) and Xiong *et al.* (2009), or the heavily lexicalized ITGs typically induced by the binarization method of Zhang

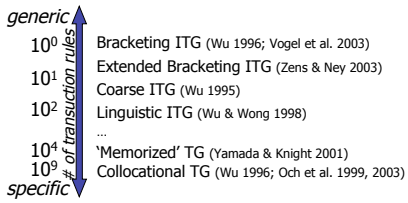


Fig. 4. Spectrum of generic to specific transduction grammars.

[1] Just as any SDTG (or synchronous CFG) that is right regular is a finite-state transduction grammar.

[2] As for example is done in ISI systems (Galley *et al.*, 2006).

[3] Subject to certain conditions; see Wu and Fung (2005) and (Wu, 1997).

et al. (2006) or the hierarchical phrase-based translation method of Chiang (2007)?

- *Headedness?* Is performance ultimately significantly improved by marking heads on the rules of an ITG to yield head ITG models—for example, as in the dependency-oriented variants of Alshawi *et al.* (1998), Cherry and Lin (2003), and Cherry and Lin (2007)?
- *Bias toward input or output language?* Bilingual transduction grammars inherently suffer some degree of mismatch between the grammar of the input versus output languages. Biasing them toward input grammars parses the input more accurately, but biasing toward output grammars as in Wu and Wong (1998) ensures grammatical translations. Toward which direction is it better to be biased, and in what way?
- *Better methods for inducing ITGs?* ITGs can be induced in many ways. The hierarchical phrase-based translation method of Chiang (2007) is one way to learn one style of ITGs. Another method is to approximate SDTGs (synchronous CFGs) with ITGs produced via the binarization algorithm of Zhang *et al.* (2006). A large space of other approaches remain to be explored.
- *How can the mismatch between training and testing conditions be reduced?* An overwhelming proportion of segment-based and tree-based SMT models still bootstrap training via relatively poor quality token-based IBM alignments, that bear little resemblance to the models used by the decoder. Recent work by Saers and Wu (2009) and Saers *et al.* (2009) begins to address this by replacing the IBM alignments with EM-trained BITG alignments right from the start, yielding improved overall translation accuracy.

Speaking more generally, past work has been lopsidedly focused on approaches that emphasize search biases. What other language biases can be exploited? And, of course, despite our emphasis on the fact that language biases are more effective than search biases when well fit to the domain, discovering improved search biases remains as important as ever.

V. Semantic SMT: Sense disambiguation

Sense disambiguation uses clues in the input context to predict the correct meaning of input lexemes that are ambiguous—thus making or influencing decisions on translation lexical choice. A large body of work on word sense disambiguation (WSD) exists, not only in the linguistic tradition, but also, more recently, in extensive empirical and machine learning oriented evaluations. Yet surprisingly, the application of semantic modeling to SMT has received little or no attention. Leveraging such modeling approaches to improve the adequacy of translation would seem highly desirable in the face of the types of semantic errors made by today's more *n*-gram based, fluency-oriented SMT systems.

Choosing the right lexical translation for an input word or phrase is essentially the same problem as choosing its sense. To accomplish this, WSD models make heavy use of the local and sentence-level context surrounding the input word—unlike current SMT models. WSD models are complementary to SMT models in this sense. However, early attempts at using context-rich approaches from Word Sense Disambiguation (WSD) methods in standard SMT systems surprisingly did not yield the expected improvements in translation quality (Carpuat and Wu, 2005).

Only recently have methods begun to emerge for successfully applying WSD models to context-dependent translation lexical choice, such as Carpuat and Wu (2007a), Chan *et al.* (2007), and Giménez and Màrquez (2007). The **phrase sense disambiguation** (PSD) model of Carpuat and Wu (2007b) leverages the feature engineering and learning models developed for standalone WSD, but succeeds where its predecessors failed by making three key adaptations:

1) *The sense disambiguation model must be trained to predict observable senses that are the direct lexical translations of the target lexeme being disambiguated.* PSD sense inventories are exactly the phrasal translations learned in the SMT translation lexicon. In contrast, most conventional WSD models are instead trained to predict hidden senses drawn from an artificially constructed sense inventory. This differs even from previous WSD approaches like those of Dagan and Itai (1994) or Gale *et al.* (1992) that make use of word translations as a source of WSD labels, but use manually-defined word-based translation lexicons rather than learned phrasal translations, and still make a distinction between sense labels and SMT translation candidates (Brown *et al.*, 1991).
2) *Sense disambiguation must be redefined to move beyond the particular case of single-token "word" targets, and instead to generalize to multi-token phrase segment targets.* PSD targets are permitted to be phrasal lexemes composed of smaller lexemes, while standalone WSD targets are typically defined as single tokens, as in Senseval tasks (e.g., Kilgarriff and Rosenzweig (1999); Kilgarriff (2001); Mihalcea *et al.* (2004)).
3) *The sense disambiguation model must be fully integrated into the runtime decoding.* Unlike earlier models attempting to utilize single-token word sense disambiguation—e.g., Carpuat and Wu (2005)—it is not possible to represent phrasal sense predictions as input annotations since they cover overlapping spans in the input sentence. PSD is fully integrated into the decoding search itself, as opposed to preprocessing or postprocessing stages.

Clearly, lessons learned from work on syntax for SMT also apply to semantics. With these three adaptations, PSD consistently yields gains across multiple Chinese-English test sets on all eight of the most commonly used automatic evaluation metrics. Work to better understand and improve sense disambiguation for SMT is ongoing; see Carpuat and Wu (2008) for more recent analysis.

VI. Semantic SMT: Semantic roles

Aside from WSD, the other major area of lexical semantics is **semantic role labeling** (SRL). Confusion of semantic roles causes translation errors that often result in serious misunderstandings of the essential meaning of the source utterances—who did what to whom, for whom or what, how, where, when, and why.

First results have begun to appear for applying shallow semantic parsing and semantic role labeling models directly to SMT, in ways that might reduce role confusion errors in the translation output (Wu and Fung, 2009b) by exploiting increasingly sophisticated models for shallow semantic parsing. Such semantic parsers, which automatically label the predicates and arguments (roles) of the various semantic frames in a sentence, are used to automatically identify inconsistent semantic frame and role mappings between the input source sentences and their output translations (Wu and Fung, 2009a). This approach is supported by the results of Fung *et al.* (2006), which reported that (for the English-Chinese language pair) approximately 84% of semantic role mappings remained consistent cross-lingually across sentence translations.

While work on semantic roles for SMT is at a nascent stage, error analyses suggest that it promises to be one of the most important directions for addressing the limitations of the current SMT paradigms.

VII. Conclusion

We have surveyed some central issues in the historical, current, and future landscape of machine translation research: a three-dimensional **MT model space**; a **socio-geographical conceptual disparity hypothesis** explaining why language pairs like Chinese-English present MT with so much difficulty; the evolution from simple **token-based** to **segment-based** to **tree-based** SMT; **language bias** rationales for selecting the degree of compositional power within the expressiveness **hierarchy of classes of transduction grammars** (or synchronous grammars), **inversion transduction grammars**, and the **ITG hypothesis**; key open questions for **syntactic SMT**; principles for successful application of **phrase sense disambiguation** models to semantic SMT; and early directions in the use of **semantic role labeling** for **semantic SMT**.

The multitude of specific problems being actively investigated in machine translation research are far too numerous to survey in the present short format. The variety may be illustrated by a few brief examples. (a) An active area of research concerns learning *without* massive parallel corpora; an obvious limitation of the currently dominant framework for SMT is its dependence on the availability of parallel texts. This is not a reasonable assumption for low-resource languages. Also, from a scientific or cognitive modeling standpoint, children learn language without the luxury of unlimited parallel text. How can SMT be learned from **small parallel corpora**? And how can SMT be learned from large amounts of monolingual **non-parallel corpora**? (b) How can multiple MT architectures be leveraged? Methods for **parallel system combination** leverage the hypothesis generation ability of different competing models. Methods for **serial system combination** use one MT model to correct the errors of another previous MT model. (c) How can **confidence estimation** techniques common in other pattern recognition applications be used to improve machine translation?

One important methodological question that requires closer scrutiny is whether current evaluation metrics are sufficiently sensitive to grammaticality improvements and semantic role improvements. If not, a vacuum of incentive will impede research.

Acknowledgement

Thanks to my numerous collaborators on these lines of work, especially Pascale Fung, Marine Carpuat, Markus Saers, Chi-kiu Lo, Yongsheng Yang, Yihai Shen, Zhaojun Wu, Hongsing Wong, Huaqing Luo, and Tyler Barth. This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under GALE Contract No. HR0011-06-C-0023, and by the Hong Kong Research Grants Council (RGC) research grants GRF621008, DAG03/04.EG09, RGC6256/00E, and RGC6083/99E. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

References

Alfred V. Aho and Jeffrey D. Ullman. Syntax-directed translations and the pushdown assembler. *Journal of Computer and System Sciences*, 3(1):37–56, 1969.

Alfred V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation, and Compiling (Volumes 1 and 2)*. Prentice-Halll, Englewood Cliffs, NJ, 1972.

Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. Automatic acquisition of hierarchical transduction models for machine translation. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 41–47, Montreal, Aug 1998.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, Robert L. Mercer, and Paul S. Roossin. A statistical approach to language translation. In *12th International Conference on Computational Linguistics (COLING-88)*, pages 71–76, Budapest, August 1988.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. Word-sense disambiguation using statistical methods. In *29th Annual Meeting of the Association for Computational Linguistics (ACL-91)*, pages 264– 270, Berkeley, Jun 1991.

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

Marine Carpuat and Dekai Wu. Word sense disambiguation vs. statistical machine translation. In *43rd Annual Meeting of*

the Association for Computational Linguistics (ACL-2005), pages 387–394, Ann Arbor, Jun 2005.

Marine Carpuat and Dekai Wu. Context-dependent phrasal translation lexicons for statistical machine translation. In *Machine Translation Summit XI (MT Summit XI)*, Copenhagen, Sep 2007.

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72, Prague, Jun 2007.

Marine Carpuat and Dekai Wu. Evaluation of context-dependent phrasal translation lexicons for statistical machine translation. In *Sixth International Conference on Language Resources and Evaluation (LREC-2008)*, Marrakech, May 2008. European Language Resources Association (ELRA).

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word sense disambiguation improves statistical machine translation. In *45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, Prague, June 2007.

Colin Cherry and Dekang Lin. A probability model to improve word alignment. In *41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 88–95, Sapporo, Aug 2003.

Colin Cherry and Dekang Lin. Inversion Transduction Grammar for joint phrasal translation modeling. In Dekai Wu and David Chiang, editors, *NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation (SSST)*, pages 17–24, Rochester, NY, April 2007.

David Chiang and Dekai Wu, editors. *Proceedings of SSST-2, Second Workshop on Syntax and Structure in Statistical Translation, at ACL-08:HLT*. Association for Computational Linguistics, Columbus, OH, Jun 2008.

David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, 2007.

Ido Dagan and Alon Itai. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596, 1994.

Pascale Fung, Zhaojun Wu, Yongsheng Yang, and Dekai Wu. Automatic learning of Chinese-English semantic structure mapping. In *IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT 2006)*, Aruba, Dec 2006.

William A. Gale, Kenneth W. Church, and David Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439, 1992.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. Scalable inference and training of context-rich syntactic translation models. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-2006*, pages 961–968, Sydney, Australia, July 2006. Association for Computational Linguistics.

Daniel Gildea. Dependencies vs. constituents for tree-based alignment. In *2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 214–221, Barcelona, Jul 2004.

Jesús Giménez and Lluís Màrquez. Context-aware discriminative phrase selection for statistical machine translations. In *Workshop on Statistical Machine Translation*, Prague, Jun 2007.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. Better word alignments with supervised ITG models. In *47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009)*, pages 923–931, Singapore, Aug 2009.

Liang Huang, Hao Zhang, and Daniel Gildea. Machine translation as lexicalized parsing with hooks. In *International Workshop on Parsing Technologies (IWPT'05)*, Vancouver, 2005.

Adam Kilgarriff and Joseph Rosenzweig. Framework and results for english senseval. *Computers and the Humanities*, 34(1):15–48, 1999. Special issue on SENSEVAL.

Adam Kilgarriff. English lexical sample task description. In *Senseval-2, Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 17–20, Toulouse, France, July 2001. SIGLEX, Association for Computational Linguistics.

Philip M. Lewis and Richard E. Stearns. Syntax-directed transduction. *Journal of the Association for Computing Machinery*, 15(3):465–488, 1968.

Rada Mihalcea, Timothy Chklovski, and Adam Killgariff. The senseval-3 english lexical sample task. In *Third International Workshop on Evaluating Word Sense Disambiguation Systems (Senseval-3)*, pages 25–28, Barcelona, Spain, July 2004. SIGLEX, Association for Computational Linguistics.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In Alick Elithorn and Ranan Banerji, editors, *Artificial and Human Intelligence: Edited Review Papers Presented at the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180. North-Holland, Amsterdam, 1984.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. Improved alignment models for statistical machine translation. In *1999 Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLPVLC-99)*, pages 20–28, College Park, MD, Jun 1999.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, Jul 2003.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translations. In *40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, pages 311–318, Philadelphia, Jul 2002.

Markus Saers and Dekai Wu. Improving phrase-based

translation via word alignments from Stochastic Inversion Transduction Grammars. In *Proceedings of SSST-3, Third Workshop on Syntax and Structure in Statistical Translation (at NAACL HLT 2009)*, pages 28– 36, Boulder, CO, Jun 2009.

Markus Saers, Joakim Nivre, and Dekai Wu. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithms. In *11th International Conference on Parsing Technologies (IWPT'09)*, Paris, Oct 2009.

Warren Weaver. Translation. In William N. Locke and A. Donald Booth, editors, *Machine Translation of Languages (1955, reprinting Weaver 1949)*, pages 15–23. MIT Press, Cambridge, MA, 1949.

Dekai Wu and David Chiang, editors. *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*. Association for Computational Linguistics, Rochester, New York, Apr 2007.

Dekai Wu and David Chiang, editors. *Proceedings of SSST-3, Third Workshop on Syntax and Structure in Statistical Translation, at NAACL-HLT 2009*. Association for Computational Linguistics, Boulder, CO, Jun 2009.

Dekai Wu and Pascale Fung. Inversion Transduction Grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Second International Joint Conference on Natural Language Processing (IJCNLP 2005)*, pages 257–268, Jeju, Korea, Oct 2005.

Dekai Wu and Pascale Fung. Can semantic role labeling improve SMT? In *13th Annual Conference of the European Association for Machine Translation (EAMT 2009)*, pages 218–225, Barcelona, May 2009.

Dekai Wu and Pascale Fung. Semantic roles for SMT: A hybrid two-pass model. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2009)*, Boulder, CO, Jun 2009.

Dekai Wu and Hongsing Wong. Machine translation with a stochastic grammatical channel. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, Montreal, Aug 1998.

Dekai Wu, Marine Carpuat, and Yihai Shen. Inversion Transduction Grammar coverage of Arabic-English word alignment for tree-structured statistical machine translation. In *IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT 2006)*, Aruba, Dec 2006.

Dekai Wu. An algorithm for simultaneously bracketing parallel texts by aligning words. In *33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, pages 244–251, Cambridge, MA, Jun 1995.

Dekai Wu. Trainable coarse bilingual grammars for parallel text bracketing. In *Third Annual Workshop on Very Large Corpora (WVLC-3)*, pages 69– 81, Cambridge, MA, Jun 1995.

Dekai Wu. Stochastic Inversion Transduction Grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404, Sep 1997.

Dekai Wu. MT model space: Statistical vs. compositional vs. example-based machine translation. *Machine Translation*, 19:213–227, 2005.

Dekai Wu. What on earth can be done about Chinese MT? (panel slides from `Chinese MT'). In *DARPA GALE meeting*, Tampa, FL, Apr 2008.

Deyi Xiong, Qun Liu, and Shouxun Lin. Maximum entropy based phrase reordering model for statistical machine translation. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-2006*, pages 521– 528, Sydney, Australia, July 2006. Association for Computational Linguistics.

Deyi Xiong, Min Zhang, Aiti Aw, and Haizhou Li. A source dependency model for statistical machine translation. In *Twelfth Machine Translation (MT Summit XII)*, pages 371–378, Ottawa, Aug 2009.

Richard Zens and Hermann Ney. A comparative study on reordering constraints in statistical machine translation. In *41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 192– 202, Sapporo, Aug 2003.

Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. Reordering constraints for phrase-based statistical machine translation. In *20th International Conference on Computational Linguistics (COLING-04)*, Geneva, Aug 2004.

Hao Zhang and Daniel Gildea. Stochastic lexicalized inversion transduction grammar for alignment. In *43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pages 475–482, Ann Arbor, Jun 2005.

Hao Zhang and Daniel Gildea. Inducing word alignments with bilexical synchronous trees. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ ACL-2006*, pages 953– 960, Sydney, Australia, July 2006. Association for Computational Linguistics.

Hao Zhang, Liang Huang, Daniel Gildea, and Kevin Knight. Synchronous binarization for machine translation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2006)*, pages 256–263, New York, June 2006. Association for Computational Linguistics.