Supertags as Source Language Context in Hierarchical Phrase-Based SMT

Rejwanul Haque, Sudip Kumar Naskar, Antal van den Bosch[†] and Andy Way

CNGL, School of Computing
Dublin City University, Dublin 9, Ireland
{ *rhaque, snaskar, away* }*@computing.dcu.ie*
[†]ILK Research Group, Tilburg centre for Creative Computing,
Tilburg University, Tilburg, The Netherlands
Antal.vdnBosch@uvt.nl

Abstract

Statistical machine translation (SMT) models have recently begun to include source context modeling, under the assumption that the proper lexical choice of the translation for an ambiguous word can be determined from the context in which it appears. Various types of lexical and syntactic features have been explored as effective source context to improve phrase selection in SMT. In the present work, we introduce lexico-syntactic descriptions in the form of supertags as source-side context features in the state-of-the-art hierarchical phrase-based SMT (HPB) model. These features enable us to exploit *source similarity* in addition to *target similarity*, as modelled by the language model. In our experiments two kinds of supertags are employed: those from lexicalized tree-adjoining grammar (LTAG) and combinatory categorial grammar (CCG). We use a memory-based classification framework that enables the efficient estimation of these features. Despite the differences between the two supertagging approaches, they give similar improvements. We evaluate the performance of our approach on an English-to-Dutch translation task, and report statistically significant improvements of 4.48% and 6.3% BLEU scores in translation quality when adding CCG and LTAG supertags, respectively, as context-informed features.

1 Introduction

The state-of-the-art hierarchical phrase-based SMT model (Chiang, 2007) uses the bilingual phrase pairs of phrase-based SMT (PBSMT) (Koehn et al., 2003) as a starting point to learn hierarchial rules using probabilistic synchronous context-free grammar (PSCFG). The decoding process in the hierarchical phrase-based SMT (HPB) model is based on bottom-up chart parsing (Chiang, 2007). This chart parsing decoder, also known as Hiero, does not require explicit syntactic representation on either side of the phrases in rules.

State-of-the-art SMT models (Koehn et al., 2003; Chiang, 2007) can be viewed as log-linear combinations of features (Och and Ney, 2002) that usually comprise translational features and the language model. The translational features typically involved in these models express dependencies between the source and target phrases, but not dependencies between the phrases in the source language themselves, i.e. they do not take into account the contexts of those phrases.

Word sense disambiguation (WSD), a task intricately related to MT, typically employs rich context-sensitive features to determine contextually the most likely sense of a polysemous word. Inspired by these context-rich WSD techniques, researchers have tried to integrate various contextual knowledge sources into state-of-the-art SMT models. In recent years, source context modelling has been successfully employed in PBSMT by taking various contextual information of the source phrase into account. These contextual features may include lexical features of words appearing in the context and bearing sense discriminatory information, position-specific neighbouring words (Giménez and Màrquez, 2007; Stroppa et al., 2007), shallow and deep syntactic

features (Gimpel and Smith, 2008), full sentential context (Carpuat and Wu, 2007), lexical syntactic descriptions in the form of supertags (Haque et al., 2009a) and grammatical dependency relations (Haque et al., 2009b).

A limitation that Hiero (Chiang, 2007) shares with the PBSMT model (Koehn et al., 2003) is that it does not take into account the contexts in which the source-sides of the rules appear. In other words, it can be argued that rule selection in Hiero is suboptimally modelled. So far, a small number of studies have made use of source-language context for improving rule selection in Hiero. Position-specific neighbouring words and their part-of-speech (POS) prove to be effective source contexts in the HPB model (He et al., 2008). In a study involving PBSMT, Haque et al. (2009b) showed that the translations of ambiguous words are also influenced by more distant words in the sentence. Syntactic contexts that capture long-distance dependencies between words in a sentence can be a useful means to disambiguate among translations. Accordingly, integration of such syntactic contexts could lead to improved translation quality in PBSMT. For instance, Haque et al. (2009a) showed that supertags are more powerful source contexts than neighbouring words and part-of-speech tags to disambiguate a source phrase in PBSMT.

Inspired by (Haque et al., 2009a), in the present work we extend the state-of-the-art Hiero system by adopting its lexical entries with the robust and efficient supertagging approaches. Grammars in these approaches consist of a syntactically rich lexicon and a small set of combinatory operators. These combinatory rules combine syntactically rich lexical entries together to form parse trees. Supertaggers assign a syntactic structure (an elementary tree or a lexical category) to each word in a sentence. These syntactic structures ('supertag') provide rich and complex linguistic information that describe the POS tag of a word, its subcategorisation information, and the hierarchy of phrase categories in which the word appears.

The remainder of the paper is organized as follows. In Section 2 we discuss related work. Section 3 provides a brief overview of HPB. In Section 4 we describe the context-informed features contained in our baseline HPB model. In Section 5 we de-scribe our memory-based classification approach. Section 6 describes experimental set-ups. Section 7 presents the results obtained, and offers a brief qualitative analysis. In Section 8 we formulate our conclusions, and offer some avenues for further work.

2   Related Work

MT research on incorporating contexts into SMT models can be broadly divided into two categories: source-context modelling such as (Stroppa et al., 2007), and target-context modelling such as (Berger et al., 1996; Hasan et al., 2008). The present study relates to the first category, which further divides into the following approaches:

Discriminative word alignment: García-Varea et al. (2001) present a MaxEnt approach to integrate contextual dependencies into the EM algorithm of the statistical alignment model to develop a refined context-dependent lexicon model. Subsequently, more recent discriminative approaches employ source-side contexts for creating finer-grained word-to-word lexicons (Brunning et al., 2009; Mauser et al., 2009; Patry and Langlais, 2009).

Phrase-based SMT: Vickrey et al. (2005) build WSD-inspired classifiers to fill in blanks in partially completed translations. Stroppa et al. (2007) were the first to add source-side contextual features into a state-of-the-art log-linear PBSMT system by incorporating context-dependent phrasal translation probabilities learned using a decision-tree classifier (Daelemans and van den Bosch, 2005). Significant improvements over a baseline PBSMT system were obtained on Italian-to-English and Chinese-to-English IWSLT tasks. Discriminative learning approaches in SMT such as (Cowan et al., 2006) generally require a redefinition of the training procedure; in contrast, Stroppa et al. (2007) introduce new features while retaining the strength of existing state-of-the-art systems. Other recent approaches to integrate state-of-the-art WSD methods into PBSMT (Giménez and Màrquez, 2007; Carpuat and Wu, 2007) have met with success as well. Following the work of (Stroppa et al., 2007), rich and complex syntactic structures such as supertags (Haque et al., 2009a) and grammatical dependency relations (Haque et al., 2009b) have been modelled as useful source context to improve phrase selection in PB-

SMT.

Alternative SMT architectures: Bangalore et al. (2007) propose an SMT architecture based on stochastic finite state transducers, that addresses global lexical selection in which parameters are discriminatively trained using a MaxEnt model considering n-gram features from the source sentence. Specia et al. (2008) integrate WSD predictions for the reranking of n-best translations, limited to a small set of words from different grammatical categories. Gimpel and Smith (2009) present an MT framework based on lattice parsing with a quasi-synchronous grammar that can incorporate arbitrary features from both source and target sentences.

Hierarchial phrase-based SMT: Chan et al. (2007) were the first to use a WSD system to integrate additional features in the state-of-the-art HPB system (Chiang, 2007), achieving statistically significant performance improvements for several automatic measures for Chinese-to-English translation. However, they only focused on solving ambiguities for those Chinese phrases that consist of only one or two terminal symbols. More recently, Shen et al. (2009) proposed a method to include linguistic and contextual information in the HPB system. The features employed in the system are non-terminal labels, non-terminal length distribution, source context and a language model created from source-side grammatical dependency structures. While their source-side dependency language model does not produce any improvement, the other features seem to be effective in Arabic-to-English and Chinese-to-English translation. Chiang et al. (2009) define new translational features using neighbouring word contexts of the source phrase, which are directly integrated into the translation model of Hiero system. In order to limit the the size of their model, they restrict words to being among the 100 most frequently occurring words from the training data; all other words are replaced with a special token. One final paper in this strand of research is that of (He et al., 2008), who despite not mentioning the link between the two pieces of work, show that the low-level source-language features used by (Stroppa et al., 2007) are also of benefit when used with the HPB decoder (Chiang, 2007).

In this paper, we present a novel approach to integrating lexical syntactic descriptions in the form of supertags as new contextual features in the HPB model. Analogous to (Stroppa et al., 2007), we use a memory-based classification approach (Daelemans and van den Bosch, 2005) to obtain probabilities for rules on the basis of additional contexts at the source-side of these rules. Some interesting properties of such classifiers include: (a) training can be performed efficiently, even with millions of examples, (b) any number of output classes can be handled, (c) the output can be seen as a posterior distribution.

3 Baseline Model

The Hierarchical PBSMT model (Chiang, 2007) is based on PSCFG. Synchronous rules in Hiero take the form as in (1):

$$X \to \langle \alpha, \gamma, \sim \rangle \qquad (1)$$

where X is the nonterminal (NT) symbol, and α and γ are the source and target phrases, which contain combinations of terminal and nonterminals in the source and target language. The $\sim$ symbol indicate a one-to-one correspondence between NTs in α and γ. In practice, the number of NTs on the right hand side is constrained to at most two, which must be separated by lexical items in α.

Each rule is associated with a score that is derived using the log-linear model (Och and Ney, 2002) as in (2):

$$w(X \to \langle \alpha, \gamma, \sim \rangle) = \sum_i \lambda_i \phi_i \qquad (2)$$

where ϕ_i is a feature defined on rules and λ_i is the feature weight of ϕ_i. One intuitively natural feature is phrase translation log-probability $\phi(\alpha, \gamma)= \log P(\gamma|\alpha)$. The other typical features used in Hiero are derived from the inverse phrase translation probability $P(\alpha|\gamma)$, the lexical probability $P_{\text{lex}}(\gamma|\alpha)$ and its inverse $P_{\text{lex}}(\alpha|\gamma)$. In the hierarchical model, translation probabilities are estimated using a relative frequency count for a phrase pair $\langle \alpha, \gamma \rangle$ independent of any other context information. Our context-informed model will be expressed as an additional feature in the model. In addition to these features the system generally employs a word penalty, a phrase penalty, a glue rule penalty, and language model features. The translation task

in HPB can be expressed as a CKY parsing with beam search together with a post-processor for mapping source derivations to target derivations (Chiang, 2007).

4 Context-informed Features

Dependencies between the consecutive source phrases (α) are not directly expressed in the HPB model. However, a discriminative classification approach to MT can be used to take into account relevant dependencies among the source phrases. Disambiguation sub-problems in MT can be partly tackled by using the direct context of the entity to be disambiguated. In other words, context-informed phrase translation can be expressed as a multi-class classification problem, where a source phrase with given additional context information is classified into a distribution over possible target phrases. This distribution may be considerably smaller than the set of possible translations of the source phrase regardless of context.

A context-informed feature ϕ_{mbl} can be viewed as the conditional probability of the target phrases γ given the source phrase α and its context information (CI), which is expressed as in (3):

$$\phi_{\mathrm{mbl}}(\alpha, \gamma) = \log \mathrm{P}(\gamma | \alpha, \mathrm{CI}(\alpha)) \qquad (3)$$

Here, CI may include any feature (lexical, syntactic, etc.), which can provide useful information to disambiguate the given source phrase. The lexical and syntactic features used in our experiments are described in the following subsections.

4.1 Lexical Feature

These features include the direct left and right context words of length i (resp. $\mathrm{w}_{\alpha-i}, ..., \mathrm{w}_{\alpha-1}$ and $\mathrm{w}_{\alpha+1}, ..., \mathrm{w}_{\alpha+i}$) of a given source phrase α. In our experiments, we consider a context size of 2 (i.e., $i := 2$). It also includes boundary words ($\mathrm{w}_{\mathrm{nt}_j^{\mathrm{start}}}$ and $\mathrm{w}_{\mathrm{nt}_j^{\mathrm{end}}}$) of subphrases covered by nonterminals in the α. Like (Chiang, 2007), we restrict the number of nonterminals to two (i.e., $j := 2$). The resultant lexical features form a window of size $2(i+j)$ features. Thus, lexical contextual information ($\mathrm{CI}_{\mathrm{lex}}$) can be described as in (4):

$$\begin{aligned} \mathrm{CI}_{\mathrm{lex}}(\alpha) = \quad & \{\mathrm{w}_{\alpha-i}, ..., \mathrm{w}_{\alpha-1}, \mathrm{w}_{\alpha+1}, ..., \mathrm{w}_{\alpha+i}, \\ & \mathrm{w}_{\mathrm{nt}_1^{\mathrm{start}}}, \mathrm{w}_{\mathrm{nt}_1^{\mathrm{end}}}, ..., \mathrm{w}_{\mathrm{nt}_j^{\mathrm{start}}}, \mathrm{w}_{\mathrm{nt}_j^{\mathrm{end}}}\} \end{aligned}$$
$$(4)$$

4.2 Syntactic Features

4.2.1 Part-of-Speech tag

In addition to the lexical features, it is possible to exploit several knowledge sources characterizing the context. For example, we can consider the POS of each word in the lexical features in (4). Contextual information ($\mathrm{CI}_{\mathrm{pos}}$) defining POS features is described as in (5):

$$\mathrm{CI}_{\mathrm{pos}}(\alpha) = \{\mathrm{pos}(w_k)\} \qquad (5)$$

where $\forall k \in [1, |\mathrm{CI}_{\mathrm{lex}}|] : w_k \in \mathrm{CI}_{\mathrm{lex}}$.

4.2.2 Supertags

Besides using local words and POS-tags as features, we introduce supertags as a syntactic source context feature in the HPB model, as in (He et al., 2008). Supertags (see Figure 1) represent complex linguistic categories that express the specific local behaviour of a word in terms of the arguments it takes (e.g. subject, object) and the syntactic environment in which it appears.
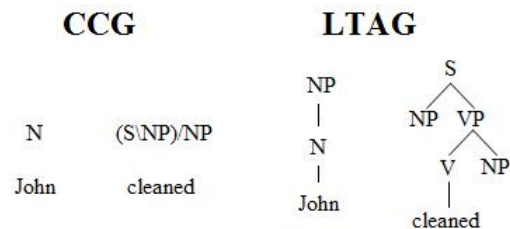

Figure 1: CCG and LTAG supertag sequences.

In our experiments two kinds of supertags are employed: those from lexicalized tree-adjoining grammar, LTAG (Joshi and Schabes, 1992), and combinatory categorial grammar, CCG (Steedman, 2000). Both the LTAG and the CCG supertag sets were acquired from the WSJ section of the Penn-II Treebank using hand-built extraction rules. Here we use both the LTAG and CCG supertaggers. In LTAG, a lexical item is associated with an elementary tree,

while in CCG the supertag constitutes a CCG lexical category with a set of word-to-word dependencies. The two alternative supertag descriptions can be viewed as closely related functional descriptors of words. Like CI_{pos}, we define the contextual information (CI_{st}) defining supertags as in (6):

$$\text{CI}_{\text{st}}(\alpha) = \{\text{st}(w_k)\} \qquad (6)$$

where $\forall k \in [1, |\text{CI}_{\text{lex}}|] : w_k \in \text{CI}_{\text{lex}}$.

Similar to the CI_{lex} feature, the syntactic features form a window of size $2(i + j)$. We compare the effect of supertag features in contrastive experiments using words and POS tags as context in order to observe the relative effects of different features. In addition, we combine the syntactic features with the lexical features. For instance, when supertags are combined with lexical features, the CI is formed by the union of these features, i.e., $\text{CI} = \text{CI}_{\text{st}} \bigcup \text{CI}_{\text{lex}}$.

5 Memory-Based Disambiguation

As Stroppa et al. (2007) point out, directly estimating context-dependent phrase translation probabilities using relative frequencies is problematic. Indeed, (Zens and Ney, 2004) showed that the estimation of phrase translation probabilities using relative frequencies results in overestimation of the probabilities of long phrases. Accordingly, smoothing factors in the form of lexical-based features are often used to counteract this bias (Foster et al., 2006). In the case of context-informed features, since the context is also taken into account, this estimation problem can only become worse.

As an alternative, in this work we make use of memory-based machine learning classifiers able to estimate $\text{P}(\gamma | \alpha, \text{CI}(\alpha))$ by similarity-based reasoning over memorized nearest-neighbour examples of source–target phrase translations, matched to a new source phrase to be translated. In this work we use the approximate memory-based classifier IGTree[1] (Daelemans and van den Bosch, 2005).

IGTree makes a heuristic approximation of k-nearest neighbour search by storing examples of source-target translation instances in the form of lossless-compressed decision trees, and performing

a top-down traversal of this tree (Daelemans et al., 1997). IGTree preserves the labeling information of all examples; in our case, a labeled example is a fixed-length feature-value vector representing the source phrase (as an atomic feature: both single-word and multi-word source phrases are treated as single values) and its contextual information, associated with a symbolic class label representing the associated target phrase. Gain ratio (GR) is used to determine the order in which features are tested in the tree. Prediction in IGTree is a straightforward traversal of the decision tree from the root node down, where a step is triggered by an exact match between a feature value of the new example and an arc fanning out of the current node. When the next step ends in a leaf node, the homogeneous class (i.e. a single phrase translation) stored at that node is returned; when no match is found with an arc fanning out of the current node, the distribution of possible class labels at the current node is returned; in our case, a weighted distribution of target phrase translations, where the weights denote the counts in the subset of the training set represented at the current node. The source phrase itself is intuitively the feature with the highest prediction power; it should take precedence in the similarity-based reasoning, and in fact it does, as it receives the highest GR value. In case of an input that mismatches on the source phrase, the overall target phrase distribution in the training set is returned.

6 Experimental Set-Up

6.1 Features Used

The output of memory-based classification is a set of weighted class labels, representing the possible target phrases (γ) given a source phrase (α) and its context information (CI). Once normalized, these weights can be seen as the posterior probabilities of the target phrases (γ) which thus give access to $\text{P}(\gamma | \alpha, \text{CI}(\alpha))$. Thus, from the classifier's output we can derive the feature ϕ_{mbl} defined in equation (4). In addition to ϕ_{mbl}, we derive a simple binary feature ϕ_{best}, defined as in (7):

$$\phi_{\text{best}} = \begin{cases} 1 & \text{if } \gamma \text{ maximizes } \text{P}(\gamma | \alpha, \text{CI}(\alpha)) \\ \approx 0 & \text{otherwise} \end{cases}$$

$$(7)$$

where ϕ_{best} is set to 1 when γ is one of the target phrases with highest probability according to $\mathrm{P}(\gamma|\alpha, \mathrm{CI}(\alpha))$; otherwise, ϕ_{best} is set to approximately 0.

We performed experiments by integrating these two features ϕ_{mbl} and ϕ_{best} directly into the log-linear model of Hiero. Their weights are optimized using minimum error-rate training (MERT)[2] on a held-out development set for each of the experiments.

6.2 Preprocessing

We used the open-source tree-based translation system moses-chart[3] to perform the experiments. HPB decoders such as moses-chart rely on a static rule table, represented as a list of aligned phrases accompanied by several estimated metrics. Since these features do not express the context information in which those rules occur, no context information is kept in the rule table, and there is no way to recover this information from the rule table. Like (Haque et al., 2009a), in order to take into account the context-informed features within such a decoder, we implemented a calling framework to translate the test set or development set.

7 Results and Analysis

Our intention to use supertags as a source-side contextual feature forced us to choose English as the source language, given that supertaggers for English are readily available. Experiments were carried out on the Dutch-to-English Open Subtitles corpus, which is collected as part of the Opus collection of freely available parallel corpora.[4] The corpus contains user-contributed translations of movie subtitles. The training text contains 285,321 sentences; the development set and test set each contain 1,000 sentences. Although our main focus was to observe the effect of incorporating supertags as a source contextual feature on translation quality, we also carried out experiments with different contextual features (both individually and in combination).

7.1 Automatic Evaluation

Translations generated by the systems are evaluated through commonly used automatic evaluation metrics: BLEU,[5] METEOR[6] and TER.[7] Additionally we performed statistical significance tests using bootstrap resampling methods on BLEU and METEOR (Koehn, 2004). The confidence level (%) of the improvements obtained by the context-informed systems with respect to the HPB moses-chart baseline are reported in the result tables below. For completeness, we note that moses-chart performs slightly better than the PBSMT decoder Moses,[8] as can be observed in Table 1.

The results obtained with the individual context indicators, compared to the baseline, are shown in Table 1. Metric-wise individual best scores are shown in bold. We observe that the English-to-Dutch subtitle translation task benefits from the addition of source-language context features, as the inclusion of any type of contextual feature improves upon the baseline (moses-chart) across all evaluation metrics. Adding words as source contexts adds 0.41 BLEU points (a relative improvement of 1.87%). Somewhat higher improvements are observed with the addition of POS context (0.47 BLEU; 2.15% relative increase), CCG supertags (0.73 BLEU; 3.33%), and LTAG supertags (0.63 BLEU; 2.88%). Among the individual contextual features, CCG produces the highest BLEU improvements over the baseline. However, none of the improvements are statistically significant (Koehn, 2004).

Exp.	BLEU	METEOR	TER
Moses	21.74	42.79	56.58
Baseline	*21.92*	*43.06*	*56.72*
Word	22.33 (77%)	43.23 (96%)	56.36
POS	22.39 (80%)	43.66 (96.2%)	56.68
CCG	**22.65** (90.3%)	43.83 (99.4%)	**56.27**
LTAG	22.55 (91.1%)	**43.99** (99.1%)	56.47

Table 1: Experimental results with individual features, compared against Moses and the moses-chart baseline.

When focusing on the METEOR evaluation met-

ric, we see that among the individual features, LTAG produces the highest improvements (0.93 points; 2.16% relative increase) over the baseline. Improvements in the METEOR metric are also observed for the CCG (0.77 METEOR; 1.79% relative increase), POS (0.60 METEOR; 1.4%) and Word (0.17 METEOR; 0.4%) features. Contrary to the BLEU comparisons, all the METEOR improvements with respect to the baseline are statistically significant. TER is an error metric, so lower scores indicate better performance. Improvements in TER for Word (a reduction of 0.36 TER points), CCG (0.45 TER points) and LTAG (0.25 TER points) features are quite reasonable and comparable to improvements in METEOR and BLEU evaluation metrics, except for the POS feature, which produces only 0.04 TER point reduction over the moses-chart baseline.

Subsequently, we performed experiments in which we combined the lexical features with the syntactic features. The results of these experiments are shown in Table 2. Metric-wise individual best scores are shown in bold. Combining LTAG supertags with Word features causes system performance to improve to 23.30 BLEU score, 1.38 points (a relative improvement of 6.3%) over the HPB baseline. CCG supertags combined with Word features produces an improvement of 0.98 absolute BLEU (4.48% relative increase). Improvements on both combinations are statistically significant at 99.5% and 95.1% levels of confidence, respectively. Interestingly, POS features together with word contexts cause system performance to deteriorate compared to the individual results; we observe only a 0.30 BLEU point improvement (1.38% relative increase, not statistically significant) over the baseline. Furthermore, we combine lexical features with two types of supertags (Word+CCG+LTAG), which gives a statistically significant 1.08 BLEU points improvement (4.93% relative increase) over the baseline.

The METEOR evaluation scores show similar trends for the combined set-ups. The best METEOR score (an improvement of 1.02 METEOR points; 2.37% relative increase) is obtained when words are combined with LTAG supertags. Moderate improvements over the baseline in METEOR are observed when Word+CCG, Word+POS and Word+CCG+LTAG are used. The improvements on

Exp.	BLEU	METEOR	TER
Baseline	*21.92*	*43.06*	*56.72*
Word+POS	22.22 (30%)	43.85 (93.4%)	56.93
Word+CCG	22.90 (95.1%)	44.00 (98.2%)	56.12
Word+LTAG	**23.30** (99.5%)	**44.08** (99.6%)	56.37
Word+CCG+LTAG	23.00 (99.8%)	43.89 (98.5%)	**55.87**

Table 2: Experimental results with combined features.

Word+CCG, Word+LTAG and Word+CCG+LTAG with respect to the baseline are statistically significant in terms of METEOR, while improvement on Word+POS is not.

On the TER evaluation metric, the best-performing combination, Word+CCG+LTAG, yields an absolute reduction of 0.85 TER points over the Hiero baseline. Reductions of 0.35 and 0.60 TER points over the baseline are seen with the Word+CCG and Word+LTAG combinations, respectively; the Word+POS combination again does not show any improvement.

7.2 Translation Analysis

We performed a manual qualitative analysis of differences between the translations produced by our best-performing context-informed system (Word+LTAG) and those by the Hiero baseline. Among the 1,000 test sentences, the Word+LTAG system attains a higher BLEU score than the baseline in 56 sentences, among which in 32 cases the improvement is due to a better lexical choice. The Word+LTAG system generates a more fluent output in 17 sentences. These two types of improvements overlap in 10 occasions (i.e. in 10 sentences, the improvement involves both better lexical choice and better fluency). The following are two such translation examples which show how our context-informed system improves over the baseline:

(1) input: i appreciate your help .
reference: ik waardeer je hulp .
Word+LTAG: ik waardeer je hulp .
baseline: ik waardeer je helpen .

(2) input: we' re taking the girl now .
reference: we halen het meisje nu .
Word+LTAG: we nemen het meisje nu .
baseline: nemen we de meisje nu .

In the first example, the word 'help' in the source English sentence is ambiguous as it can translate to the noun 'hulp' or the verb 'helpen'. The Word+LTAG system conveys a meaning more similar to the input sentence by choosing the correct 'hulp'. In the second example, the translation of the Word+LTAG system is more fluent than the baseline Hiero translation, as it generates a correct word order while the baseline does not, and it chooses the correct neuter article 'het' instead of the incorrect non-neuter article 'de' selected by the baseline.

As an additional analysis, we examined the decoding process to discover why the Word+LTAG system generates better output than the baseline. In the first example, to translate the source sentence, 5,354 candidate phrases are used by the baseline system, while only 460 candidate phrases (IGTree classes) are used by the Word+LTAG system. As a result, during decoding, 9,654 hypotheses are generated in the Word+LTAG system compared to 20,371 hypotheses in the baseline. We also identified details regarding what candidate phrases along with source spans are used for the best translation hypothesis. A source span for each candidate phrase is represented by word positions in the source sentence ([1..n]; where, n: sentence length). In the Word+LTAG system, candidate phrases used in the best translation hypothesis are: 'ik':[1..1], 'waardeer':[2..2], 'je hulp':[3..4] and '.':[5..5]. On the other hand, the baseline uses two candidate phrases ('ik waardeer je':[1..3] and 'helpen .':[4..5]) to generate the best translation hypothesis, and the usage of the last phrase ('helpen .') in this translation is incorrect.

In the second example, to translate the source sentence, 8,518 candidate phrases are used by the baseline system, while only 1,577 candidate phrases are used by the Word+LTAG system. As a result, during decoding, 24,092 hypotheses are generated in the Word+LTAG system compared to 35,659 hypotheses in the baseline. In the Word+LTAG system, the candidate phrases used to generate the best translation hypothesis are: 'we nemen':[1..2], 'het meisje':[2..4], 'nu':[5..5] and '.':[6..6]. By contrary, the baseline uses the following candidate phrases: 'nemen we de':[1..3], 'meisje':[4..4], 'nu':[5..5] and '.':[6..6]. The baseline system chooses an incorrect candidate phrase ('nemen we de') to generate the best translation hypothesis.

The above analysis reveals that in addition to the context-dependent translation features, context-informed models use reduced but more fine-grained sets of candidate phrases, which in turn force the model to weed out bad hypotheses during decoding, and thereby improve translation quality.

7.3 Numbers of rules and examples

Hiero usually generates a massive number of rules compared to the phrase-based approach. The first data row in Table 3 shows that the number of distinct rules (rule table size) generated by Hiero for our English-to-Dutch dataset is almost three times larger than the number of distinct source-target phrase-pairs (phrase table size) generated by Moses on the same dataset. The bottom row in Table 3 shows a similar trend in the case of all rules (non-distinct) generated from the parallel training data during the rule extraction process of Hiero. IGTree classifiers are built on the set of examples formed by the source phrase (α), target phrase (γ), and the context information (CI) of the source phrase obtained during the rule extraction process in Hiero. In other words, the number of training examples equals the number of times Hiero's rules apply to the training source sentences. Although IGTree scales roughly linearly to larger numbers of examples, it would be a challenge on present-day computers to train IGTree with one order of magnitude more data.

	Hiero	Moses
Distinct	6,761,376	1,988,504
Non-distinct	11,603,617	3,817,252

Table 3: Numbers of rules in Hiero or phrase-pairs in Moses.

8 Conclusions and Future Work

In this paper we demonstrated that supertags can be succesfully integrated as source-side contextual features into the state-of-the-art hierarchical phrase-based SMT system, Hiero. Following earlier work, we compared the integration of supertag features to the integration of contextual words and POS tags in the Hiero system. Considering only individual contextual features, the system produces better gains for supertags (with 3.33% and 2.88% relative gains in BLEU for CCG and LTAG respectively) than words

(1.87% relative gain) and POS tags (2.15% relative gain) in an English-to-Dutch translation task. Furthermore, we observed the best improvement over the baseline when combining supertags with word contexts (4.48%, 6.3% and 4.93% relative improvements in BLEU for Word+CCG, Word+LTAG and Word+CCG+LTAG respectively), while POS features together with word contexts only showed a 1.38% relative increase.

The relative lack of effect of the combination of POS tags and supertags lies in the fact that POS information is present already in the supertags. POS tags are therefore redundant when supertags are also available. Words, on the other hand, remain relevant as they appear to contain complementary information not carried by supertags. Generalizing these results, it would be interesting in future research to compare supertags to grammatical dependency relations as context features, as already explored by Haque et al. (2009b) for PBSMT. Like dependency relations, supertags describe how a word is related to its grammatical neighbours, regardless of their position. However, while supertags may capture long-distance dependencies in an indirect way, dependency relations encode direct relations; by following a dependency relation one can, for example, directly obtain the lexical identity of the related word. We plan, therefore, to model and incorporate grammatical dependency relations as source-language contextual features into the Hiero system.

Our experiments have focused on a standard but small dataset. Despite the challenges to train classifiers with large sets of instances, we intend to further validate our conclusions by scaling up to larger datasets, and perform learning curve experiments to observe changes in the relative differences between using different types of additional source-side contextual features.

Acknowledgments

References

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Alexandre Patry and Philippe Langlais. 2009. Prediction of words in statistical machine translation using a multilayer perceptron. In *Proceedings of the twelfth Machine Translation Summit (MT Summit XII)*, Ottawa, ON, Canada, pages 101–111.

Arne Mauser, Saša Hasan, and Hermann Ney. 2009. Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, (EMNLP'09)*, Singapore, pages 210–218.

Aravind K. Joshi and Yves Schabes. 1992. Tree Adjoining Grammars and Lexicalized Grammars. In M. Nivat and A. Podelski (eds.) *Tree Automata and Languages*, Amsterdam, The Netherlands: North-Holland, pages 409-431.

Brooke Cowan, Ivona Kucerov'a, and Michael Collins. 2006. A discriminative model for tree-to-tree translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, Sydney, Australia, pages 232–241.

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

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proceedings of the 2009 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'09)*, Boulder, CO, pages 218–226.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*, Vancouver, BC, Canada, pages 771–778.

Franz J. Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, PA, pages 295–302.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable smoothing for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, Sydney, Australia, pages 53-61.

Ismael García-Varea, Franz J. Och, Hermann Ney, and Francisco Casacuberta. 2001. Refined lexicon models for statistical machine translation using a maximum entropy approach. In *39th Annual meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter of ACL (ACL-EACL'01)*, Toulouse, France, pages 204–211.

Jamie Brunning, Adrià de Gispert, and William Byrne. 2009. Context-dependent alignment models for statistical machine translation. In *Proceedings of the 2009 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'09)*, Boulder, CO, pages 101–111.

Jesús Giménez and Lluís Màrquez. 2007. Context-aware discriminative phrase selection for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation, ACL'07*, Prague, Czech Republic, pages 159–166.

Kevin Gimpel and Noah A. Smith. 2008. Using Grammatical Dependencies for Contextual Phrase Translation Disambiguation. In *Proceedings of the Third Workshop on Statistical Machine Translation, ACL'08:HLT*, Columbus, OH, pages 9–17.

Kevin Gimpel and Noah A. Smith. 2009. Feature-Rich Translation by Quasi-Syncronous Lattice Parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, Singapore, pages 219–228.

Libin Shen, Bing Zhang, Spyros Matsoukas, and Ralph Weischedel. 2009. Effective use of linguistic and contextual information for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, Singapore, pages 72–80.

Lucia Specia, Baskaran Sankaran, and Maria das Graças Volpe Nunes. 2008. n-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation. In *Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics (CICLING'08)*, Haifa, Israel, pages 399–410.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation Using Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*, Prague, Czech Republic, pages 61–72.

Mark Steedman. 2005. *The Syntactic Process*, MIT Press: Cambridge, MA.

Nicolas Stroppa, Antal van den Bosch, and Andy Way. 2007. Exploiting Source Similarity for SMT using Context-Informed Features. In *Proceedings of the 11th International Conference on Theoretical anc Methodological Issues in Machine Translation (TMI'07)*, Skövde, Sweden, pages 231–240.

Philip Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, Barcelona, Spain, pages 388–395.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP'03)*, Edmonton, AB, pages 48–54.

Rejwanul Haque, Sudip Kumar Naskar, Antal van den Bosch, and Andy Way. 2009. Dependency Relations as Source Context in Phrase-Based SMT. In *The 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC'09)*, Hong Kong, China, pages 170–179.

Rejwanul Haque, Sudip Kumar Naskar, Yanjun Ma, and Andy Way. 2009. Using Supertags as Source Language Context in SMT. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, Bercelona, Spain, pages 234–241.

Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP'04)*, Boston, MA, pages 257–264.

Saša Hasan, Juri Ganitkevitch, Hermann Ney, and Jesús Andrés-Ferrer. 2008. Triplet lexicon models for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, Honolulu, HI, pages 372–381.

Srinivas Bangalore, Patrick Haffner, and Stephan Kanthak. 2007. Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction. In *Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, pages 152–159.

Walter Daelemans and Antal van den Bosch. 2005. *Memory-Based Language Processing*, Cambridge University Press, Cambridge, UK.

Walter Daelemans, Antal van den Bosch, and A. Weijters. 1997. IGTree: using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review*, 11:407–423.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, pages 33–40.

Zhongjun He, Qu Liu, and Shouxu Lin. 2008. Improving statistical machine translation using lexicalized rule selection. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING'08)*, Manchester, UK, pages 321–328.