Mach Translat (2011) 25:239–285
DOI 10.1007/s10590-011-9100-2

Integrating source-language context into phrase-based statistical machine translation

Rejwanul Haque · Sudip Kumar Naskar · Antal van den Bosch · Andy Way

Received: 30 October 2010 / Accepted: 29 July 2011 / Published online: 7 October 2011

Abstract The translation features typically used in Phrase-Based Statistical Machine Translation (PB-SMT) model dependencies between the source and target phrases, but not among the phrases in the source language themselves. A swathe of research has demonstrated that integrating source context modelling directly into log-linear PB-SMT can positively influence the weighting and selection of target phrases, and thus improve translation quality. In this contribution we present a revised, extended account of our previous work on using a range of contextual features, including lexical features of neighbouring words, supertags, and dependency information. We add a number of novel aspects, including the use of semantic roles as new contextual features in PB-SMT, adding new language pairs, and examining the scalability of our research to larger amounts of training data. While our results are mixed across feature selections, classifier hyperparameters, language pairs, and learning curves, we observe that including contextual features of the source sentence in general produces improvements. The most significant improvements involve the integration of long-distance contextual features, such as dependency relations in combination with part-of-speech tags in Dutch-to-English subtitle translation, the combination of dependency parse and semantic role information in English-to-Dutch parliamentary debate translation, or supertag features in English-to-Chinese translation.

R. Haque · S. K. Naskar · A. Way
CNGL, School of Computing, Dublin City University, Dublin 9, Ireland

A. van den Bosch (✉)
ILK Research Group, Tilburg center for Cognition and Communication,
Tilburg University, Tilburg, The Netherlands
e-mail: Antal.vdnBosch@uvt.nl

Keywords Statistical machine translation · Phrase-based statistical machine translation · Syntax in machine translation · Translation modelling · Word alignment · Memory-based classification

1 Introduction

In log-linear phrase-based statistical machine translation (PB-SMT: Koehn et al. 2003), the probability $P(\hat{e}_k \mid \hat{f}_k)$ of a target phrase $\hat{e}_k$ given a source phrase $\hat{f}_k$ is modelled as a log-linear combination of features which typically consist of a finite set of translation features, and a language model (Och and Ney 2002). The translation features normally used in such models express dependencies between the source and target phrases, but not among phrases or tokens in the source language themselves.

Stroppa et al. (2007) observed that incorporating source-language context using neighbouring words and part-of-speech tags had the potential to improve translation quality. This has led to a whole tranche of research, of which we provide an overview in Sect. 3, which has shown that integrating source context modelling into PB-SMT can positively influence the weighting and selection of target phrases, and thus improve translation quality.

Approaches to include source-language context to help select more appropriate target phrases have partly been inspired by methods used in word-sense disambiguation (WSD), where rich contextual features are employed to determine the most likely sense of a polysemous word given that context. These contextual features may include lexical features of words appearing in the immediate context (Giménez and Màrquez 2007; Stroppa et al. 2007), shallow and deep syntactic features of the sentential context (Gimpel and Smith 2008), and full sentential context (Carpuat and Wu 2007). Studies in which syntactic features are employed have made use of part-of-speech taggers (Stroppa et al. 2007), supertaggers (Haque et al. 2009a), and shallow and deep syntactic parsers (Gimpel and Smith 2008; Haque et al. 2009b).

In prior work, we have shown that exploring local sentential context information in the form of both supertags (Haque et al. 2009a) and syntactic dependencies (Haque et al. 2009b) can be successfully integrated into a PB-SMT model. Here we provide a revised, extended account of this previous research. We add a number of novel aspects, including using semantic roles as new contextual features in PB-SMT, adding new language pairs, and examining the scalability of our research to larger amounts of training data.

Our results allow us to conclude that incorporating source-language contextual features benefits a range of different language pairs, both with English as source language (translating to Dutch, Chinese, Japanese, Hindi, Spanish, and Czech) and target language (from Dutch), on different types of data such as news articles and commentary, parliamentary debates, patents, and subtitles, according to a range of automatic evaluation measures.

The remainder of this contribution is organized as follows. Section 2 provides some motivation for work in this direction, and related work is discussed in Sect. 3. Section 4 provides a brief overview of PB-SMT, which acts as the baseline throughout the study. In Sect. 5 we describe the range of context-informed features we add to the baseline

PB-SMT model. Section 6 describes the memory-based classification approach and how we integrated the output of the memory-based classifier into a state-of-the-art PB-SMT system. In Sect. 7 we present the results obtained. We formulate our conclusions in Sect. 8, and offer some avenues for further work.

2 Motivation

In PB-SMT, the selection of appropriate target phrases for a source phrase depends only on the source phrase itself, regardless of its context, and the target language model. The target language model does represent contextual information, but only of the target language. In other words, the sense-disambiguation task inherent to PB-SMT is modelled suboptimally as it ignores source-side context when translating a source phrase. We argue that the disambiguation of a source phrase can be enhanced by taking into account the context of the source phrase.

Figure 1 shows translation examples for the polysemous English word '*play*', which has many translation equivalents in Chinese, some of which ('*xi*', '*wan*', '*banYan*', '*boFang*') are shown in Fig. 1. Figure 1 also shows four English sentences, each containing '*play*', translated into four different Chinese words depending on the context. For example, the most suitable translation of the word '*play*' for the first English sentence '*He wrote a play*' is '*xi*' amongst the four Chinese candidate translations. The translation of '*play*' in this sentence depends on the neighbouring word '*wrote*'. Similarly, the appropriate translation of '*play*' in the English sentence '*Can you play my favourite old record?*' is '*boFang*' amongst the four Chinese candidate translations. In this English sentence, the translation of '*play*' depends on the distant word '*record*'.

In PB-SMT, translation of a source sentence begins by generating all possible source phrases and gathering all candidate target phrases for each source phrase. In the translation process, potentially thousands of translation hypotheses are statistically generated using the pool of all target phrases to form the candidate translations. Thus, the decoder considers all target phrases for a given source phrase as possible candidate translations of that source phrase. The candidate phrases with higher translation probabilities have a better chance of occuring in the most likely candidate translations. On the other hand, the candidate phrases with lower translation

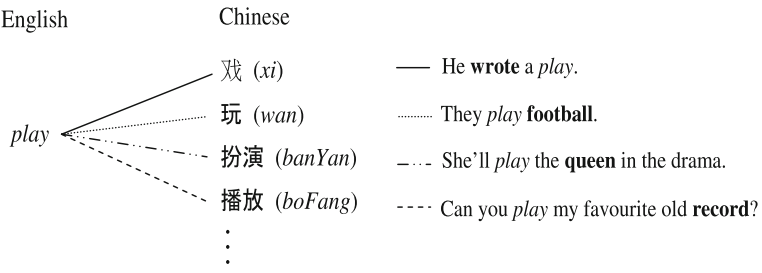

Fig. 1 Examples of ambiguity for the English word *play*, together with different translations depending on the context

probabilities have a higher risk of being pruned out during the formation of translation hypotheses due to the decoder's beam size limit. The phrase translation probabilities are measured based on the frequency of occurrences of the source and target phrase pairs in the training corpus, ignoring the context in which those phrases appear.

Let us go back to the example of the ambiguous word '*play*' in Fig. 1. Suppose the translation probability of '*play*' into '*xi*' is higher than that of '*play*' into any of the remaining Chinese words ('*wan*', '*banYan*', '*boFang*'). During translation of any of the English sentences in Fig. 1, the decoder would ignore whatever contextual dependency the source phrase '*play*' has in the source sentence, consider all Chinese words as probable candidate translations, and always give preference to '*xi*' while generating candidate translations.

To study whether we can counterbalance the important role of the target language model in the selection of candidate phrases, in this work we incorporate various aspects of source-language context into a state-of-the-art PB-SMT model, Moses (Koehn et al. 2003, 2007), in order to perform discriminative translation filtering (cf. Sect. 3.1.2) by learning context-sensitive translation probabilities which in effect should improve target phrase selection. We see from Fig. 1 that the translations of '*play*' may depend on the neighbouring lexical context ('*wrote*', in the first example sentence) as well as quite distant lexical context ('*record*', in the last example sentence). We investigate the incorporation of basic contextual features (words and POS tags), lexical syntactic descriptions (supertags), deep syntactic information (grammatical dependency relations) and semantic roles into PB-SMT. We conjecture that such kinds of complex and rich syntactic and semantic knowledge sources, some of which inherently capture long-distance word-to-word dependencies in a sentence, may be useful to improve PB-SMT lexical selection.

3 Related work

Context has been incorporated into both the source and target sides of the translation pair, in order to improve translational choice and the quality of translation. Techniques to incorporate context into SMT can be broadly divided into two categories: *source-context modelling* (Carpuat and Wu 2007; Giménez and Màrquez 2007; Stroppa et al. 2007; etc.) and *target-context modelling* (Berger et al. 1996; Hasan et al. 2008; etc.). In the first two subsections of this section we describe twenty nine studies of context modelling with English as the target language. In the third subsection we highlight six studies that use English as the source language.

In the three subsections we group related work according to six key aspects. Each aspectual overview of related work is accompanied by a table highlighting the contrastive features of the studies discussed (Tables 1, 2, 3, 4, 5 and 6). In each table we list the contextual features employed by each study and the types of SMT models employed. In the second column of each table 'SL→TL' stands for 'source language→target language', referring to the translation pair and direction; 'DS' refers to the 'data sets' used to train SMT models; and 'S/L' stands for 'small/large', indicating a division between training set sizes below and above 500,000 words that we use to structure our

Table 1 Related research integrating context into word-based SMT (WB-SMT) models

Authors	SL→TL[DS][S/L]	Contextual features	Integrated into
Brown et al. (1991)	Fr→En[CPH][L]	SL:neighbouring words and basic POS	WB-SMT model
Vickrey et al. (2005)	Fr→En[Europarl][L]	SL:POS and neighbouring words	WB-SMT model
Carpuat and Wu (2005)	Zh→En[UN:LDC][L]	SL:position-sensitive syn-tactic, and local collocations	WB-SMT model

En English, *Fr* French, *Zh* Chinese, *CPH* Canadian Parliament Hansards, *UN* United Nations, *LDC* Linguistic Data Consortium

Table 2 Related research integrating context into PB-SMT models

Authors	SL→TL[DS][S/L]	Contextual features	Integrated into
Stroppa et al. (2007)	Zh→En[IWSLT][S]	SL:neighbouring words and POS tags	PB-SMT model
	It→En[IWSLT][S]		
Carpuat and Wu (2007)	Zh→En[IWSLT][S]	SL:bag-of-words, collocations, POS and dependency features	PB-SMT model
	Zh→En[NIST][L]		
Giménez and Màrquez (2007, 2009)	Sp→En[Europarl][L]	SL:local context, *n*-grams, POS, lemmas, chunk label and bag-of-words	PB-SMT model

En English, *It* Italian, *Sp* Spanish, *Zh* Chinese

Table 3 Related research integrating context into Hiero models

Authors	SL→TL[DS][S/L]	Contextual Features	Integrated into
Chan et al. (2007)	Zh→En[FBIS][L]	SL:local collocations, POS tags and neighbouring words	Hiero
Shen et al. (2009)	Ar→En[NIST 06, 08][L]	SL:nonterminal labels, length, context LM and dependency LM	Hiero
	Zh→En[NIST 06, 08][L]		
Chiang et al. (2009)	Zh→En[NIST][L]	SL:neighbouring words	Hiero and syntax-based model

En English, *Zh* Chinese, *Ar* Arabic, *FBIS* Foreign broadcast information service

experiments. In the third column, 'SL' and 'TL' respectively stand for 'source' and 'target' languages of the translation pair from which the listed contextual features are extracted.

Table 4 Related research integrating context into alternative SMT models

Authors	SL→TL[DS][S/L]	Contextual features	Integrated into
Bangalore et al. (2007)	Ar→En[UN][L]	SL:bag-of-words	FST-based MT model
	Fr→En[CPH][L]		
	Zh→En[IWSLT][S]		
Ittycheriah et al. (2007)	Ar→En[UN, NIST 06][L]	SL:lexical, morphological and syntactic features	Proposed DTM2 model
Gimpel and Smith (2009)	De→En[BTEC][S]	SL and TL:syntactic features from dependency trees	Proposed MT model

En English, *Fr* French, *De* German, *Zh* Chinese, *Ar* Arabic, *CPH* Canadian Parliament Hansards, *UN* United Nations, *BTEC* basic travel expression corpus, *FST* finite state transducer

Table 5 Related research integrating context into word alignment models

Authors	SL→TL[DS][S/L]	Contextual features	Integrated into
Berger et al. (1996)	Fr→En[CPH][L]	TL:neighbouring words	IBM model
	De→En[Verbmobil][L]		
García-Varea et al. (2001, 2002)	De→En[Verbmobil][S]	SL and TL:neighbouring words and word class	IBM model
	Fr→En[CPH][L]		
Mauser et al. (2009)	Zh→En[GALE][L]	TL:neighbouring words and SL:sentence level lexical feature	IBM model and proposed discriminative WA model
	Ar→En[NIST 08][L]		
	Zh→En[NIST 08][L]		
Patry and Langlais (2009)	Fr→En[Europarl][S]	SL:bag-of-words	Proposed WA model

En English, *Fr* French, *De* German, *Zh* Chinese, *Ar* Arabic, *CPH* Canadian Parliament Hansards, *WA* word alignment

3.1 Source context modelling

Approaches to integrating source-language contextual information into different stages in the SMT model can in turn be broadly divided into: (i) *discriminative word alignment* (e.g. Brunning et al. 2009; Patry and Langlais 2009) for creating improved word-to-word translation lexicons, and (ii) *discriminative translation filtering* (e.g. Carpuat and Wu 2007; Chan et al. 2007; Stroppa et al. 2007) by learning context-dependent translation probabilities.

3.1.1 Discriminative word alignment

García-Varea et al. (2001, 2002) present a MaxEnt approach to integrate contextual dependencies of both the source and target sides of the statistical alignment model

Table 6 Related research using English as source language

Authors	SL→TL[DS][S/L]	Contextual features	Integrated into
Max et al. (2008)	En→Fr[Europarl][S]	SL:neighbouring words and POS, and dependency relations	PB-SMT model
Gimpel and Smith (2008)	Zh→En[NIST 08, UN][L] De→En[WMT 07][L] En→De[WMT 07, 08][L]	SL:lexical, shallow syntactic and positional features	PB-SMT model
Venkatapathy and Bangalore (2007)	En→Hi[News][S]	SL:bag-of-words	Proposed global lexical selection model
Specia et al. (2008)	En→Pt[Europarl][L]	SL:morphological features (person, tense and number)	Dependency treelet system
Hasan et al. (2008)	Zh→En[IWSLT][S] Sp→En[EPPS][L] En→Sp[EPPS][L]	TL:words	IBM model
Brunning et al. (2009)	Ar→En[NIST 08][S] En→Ar[NIST 08][S]	SL and TL:POS tag	MTTK WA model

En English, *Fr* French, *De* German, *Sp* Spanish, *Zh* Chinese, *Ar* Arabic, *Pt* Portuguese, *Hi* Hindi, *WA* Word alignment

to develop a refined context-dependent lexicon model. They report better alignment quality in terms of improved alignment error rate (AER) (Och and Ney 2000).

Patry and Langlais (2009) propose an alignment model which does not assume word alignments and considers all source words jointly when evaluating the probability of a target word. They use a multilayer perceptron classifier to estimate this probability. The word alignment results surpass IBM model 1 when their model is extended to include alignment information.

3.1.2 Discriminative translation filtering

Discriminative translation filtering in SMT, in which contextual information from the source language is used to weight or select from the potentially large set of lexical or phrasal translations, can furthermore be divided into two categories: (i) *hard interaction* (e.g. Carpuat and Wu 2005) and (ii) *soft interaction* (e.g. Chan et al. 2007; Stroppa et al. 2007). Alternatively, the same techniques can also be classified according to their use of (i) *hard constraints* (e.g. Stroppa et al. 2007) or (ii) *soft constraints* (e.g. Carpuat and Wu 2007; Marton and Resnik 2008; Xiong et al. 2010).

- *Hard vs. soft interaction*: In soft interaction, WSD-like translation predictions from context-informed translation models are allowed to interact with other log-linear models (e.g. the target language model, a distortion model, additional translation models, etc.) in the decoder. The additional context-informed model thus adds its influence to the mix of models integrated during decoding. In hard interaction, the

WSD-like translation predictions are used for pre-processing or post-processing, and do not interfere with the SMT process itself. In other words, the weights of the context-dependent translations of a source phrase are not integrated with other SMT models to select the best candidate translations.

- *Hard vs. soft constraints*: In the soft constraints model, the decoder is allowed to use all possible candidate phrases for a source phrase, while soft constraints such as weights are introduced to influence the decoder's lexical selection model. In the hard constraints model, the decoder is forced to use a restricted but supposedly more appropriate set of candidate phrases for a source phrase; in addition, the context-informed model imposes weights on the candidate phrases on the basis of additional contextual information to influence the decoder's lexical selection model.

According to this division, analogous to the work of (Stroppa et al. 2007), our context-informed models interact 'softly' with the other SMT models, and we impose hard constraints on the decoder.

Discriminative translation filtering in SMT can further be divided into the following four categories according to its deployment into different types of SMT engines:

- *Word-based SMT*: Table 1 lists related research that integrates context into word-based SMT models. Brown et al. (1991) were the first to propose the use of dedicated WSD models in word-based SMT systems, using an English-to-French translation task as their testbed. An instance of a word is assigned a sense based on mutual information with the word's translation. Evaluation is limited to the case of binary disambiguation, i.e. deciding between only the two most probable translation candidates, and to a reduced set of common words. A significant improvement in translation quality is reported, according to manual evaluation. Vickrey et al. (2005) build classifiers inspired by those used in WSD to fill in blanks in a partially completed translation. This blank-filling task is a limited subtask of the translation task, in which the (possibly incorrect) target context surrounding the word translation is already available. Carpuat and Wu (2005) integrate a WSD model into a word-based SMT system in two ways: (i) the WSD model constrains the set of potential senses considered by the decoder; and (ii) the SMT output is post-processed by directly replacing translation candidates with the WSD predictions. The integration of the WSD model into the SMT system is performed in a hard manner, and both approaches are found to hurt translation quality.
- *Phrase-Based SMT*: Table 2 summarizes related research on integrating context into PB-SMT models. Stroppa et al. (2007) successfully add source-side contextual features into a log-linear PB-SMT system (Koehn et al. 2003) by incorporating context-dependent phrasal translation probabilities learned using a decision-tree classifier (Daelemans and van den Bosch 2005). Several proposals have recently been formulated that exploit the accuracy and the flexibility of discriminative learning (Liang et al. 2006; Tillmann and Zhang 2006). Work of this type generally requires a redefinition of the training procedure; in contrast, Stroppa et al. (2007) introduce a discriminative component that can be built into PB-SMT systems, retaining the strength of the latter.

More recent approaches of integrating state-of-the-art WSD methods into PB-SMT (Carpuat and Wu 2007; Giménez and Màrquez 2007; Gimpel and Smith 2008; Max et al. 2008) have also met with success in improving the overall translation quality. Giménez and Màrquez (2007) extend the work of Vickrey et al. (2005) by (i) considering the more general case of frequent phrases and moving to full translation rather than the blank-filling task on the target side, and (ii) moving from word translation to phrase translation. Giménez and Màrquez (2009) show that their discriminative models yield significantly improved lexical choice over a PB-SMT model, which in turn does not necessarily lead to improved grammaticality. Carpuat and Wu (2007) incorporate a *phrase-sense disambiguation* model directly into a state-of-the-art PB-SMT system, producing consistent gains across all evaluation metrics on the IWSLT 2006 and NIST Chinese-to-English translation tasks.

- *Hierarchical phrase-based SMT*: Table 3 lists related research that integrates context into hierarchical PB-SMT models. Chan et al. (2007) were the first to use a WSD system to integrate additional features in hierarchical phrase-based SMT (HPB-SMT) (Chiang 2007), achieving statistically significant performance improvements for several automatic measures for Chinese-to-English translation. More recently, Shen et al. (2009) proposed a method to include linguistic and contextual information in the HPB-SMT model. While their source-side dependency language model does not produce improvements, 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 context of the source phrase, which are directly integrated into both the translation model of the Hiero system (Chiang 2007) and the syntax-based system of Galley et al. (2006).
- *Alternative SMT architectures*: Table 4 lists related research that integrates context into alternative SMT models. 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. Ittycheriah et al. (2007) introduce the Direct Translation Model 2 (DTM2) which employs discriminative MaxEnt models to obtain the translation likelihoods. 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 (Table 5).

3.2 Target context modelling

Table 6 enumerates related research that integrates context into word alignment models. Berger et al. (1996) suggest context-sensitive modelling of word translations in order to integrate local contextual information into their IBM translation models using a MaxEnt model. Probability distributions are estimated by MaxEnt based on position-sensitive local collocation features in a window of three words around the target word. This work is not supported by significant evaluation results. Mauser et al. (2009) extend the work of Hasan et al. (2008) by integrating additional discriminative word lexicons into the PB-SMT model, by using sentence-level source information to predict appropriate target words.

3.3 English as source language

It is a common belief that translating into a less inflected language (such as English) from a highly inflected language should be more effective than the other way round. This belief is hardly challenged in the related work cited in this section; all above-mentioned twenty nine studies translate to English. Nonetheless, Table 6 lists six studies that take English as the source language. Most of these studies employ contextual features computed on the English input; the ample availability of Natural Language Processing (NLP) tools for English, such as part-of-speech taggers and parsers, makes this possible. Both Max et al. (2008) and Gimpel and Smith (2008) work with a state-of-the-art PB-SMT system (Koehn et al. 2003) and focus on language pairs where the target is not English. Max et al. (2008) conduct experiments on English-to-French and show modest gains over a PB-SMT baseline model according to manual evaluation. Gimpel and Smith (2008) work with two different English-to-German data sets (WMT'07 and WMT'08) and perform a range of experiments. Most gains are not statistically significant with the WMT'07 translation task, but with the WMT'08 task most gains are statistically significant.

Venkatapathy and Bangalore (2007) conduct experiments on a small amount of English-to-Hindi training data with a global lexical selection and sentence reconstruction model. Their bag-of-words model considers all words of the source sentence as features, regardless of their positions. Using these features, a MaxEnt-based classifier predicts target words that should occur in the target sentence. The target sentence is determined by permuting the generated target words using a language model.

In an English-to-Portuguese translation task, Specia et al. (2008) work with a syntactically motivated PB-SMT system (Quirk et al. 2005), which they enrich by a WSD model limited to disambiguating ten highly frequent and ambiguous verbs.

Two more studies (Hasan et al. 2008; Brunning et al. 2009) consider English as the source language. Both approaches focus on improving word alignment for creating refined word-to-word translation lexicons. Hasan et al. (2008) present an integration of target context modelling into SMT using a triplet lexicon model that captures long-distance dependencies. Brunning et al. (2009) introduce context-dependent alignment models for MT that exploit source-language context information to estimate word-to-word translation probabilities using a decision tree algorithm.

We see from the above summary tables that a range of contextual features have been employed at different stages in the SMT model. In the present work, we integrate a range of contextual features into a state-of-the-art PB-SMT model, Moses (Koehn et al. 2003, 2007), including neighbouring words and POS tags of the source phrases as in (Giménez and Màrquez 2007; Stroppa et al. 2007). We introduce lexical syntactic descriptions in the form of supertags and semantic roles as new contextual features in the PB-SMT model. Moreover, we investigate the integration of deep syntactic information (grammatical dependency relations) into the PB-SMT models as in (Carpuat and Wu 2007; Max et al. 2008). A more detailed account of the difference between our approach and those of (Carpuat and Wu 2007; Max et al. 2008) with respect to various aspects is given in Sect. 5.4.

In this contribution we report on a range of experiments on several data sets, mostly adopting English as the source language of the translation pairs

(English-to-Hindi, English-to-Czech, English-to-Dutch, English-to-Chinese, English-to-Japanese, English-to-Spanish). We examine the scalability of our research to larger amounts of training data. Furthermore, we report on experiments with Dutch-to-English translation, using experimental data from two different domains. In the next section we provide a brief overview of our baseline PB-SMT system.

4 PB-SMT baseline

The translation task in SMT can be viewed as a search problem (Brown et al. 1993) in which the goal is to find the most probable candidate translation $e_1^I = e_1, \ldots, e_I$ for the given input sentence $f_1^J = f_1, \ldots, f_J$. The best translation can be obtained applying the noisy channel model (Brown et al. 1990) of translation by maximizing $\mathrm{P}(e_1^I | f_1^J)$, as in (1):

$$\underset{I, e_1^I}{\operatorname{argmax}}\, \mathrm{P}(e_1^I | f_1^J) = \underset{I, e_1^I}{\operatorname{argmax}}\, \mathrm{P}(f_1^J | e_1^I).P(e_1^I) \tag{1}$$

where $\mathrm{P}(f_1^J | e_1^I)$ and $\mathrm{P}(e_1^I)$ denote respectively the translation model and the target language model (Brown et al. 1993).

The log-linear translation model (Och and Ney 2002) is a special case of the noisy channel model of translation, in which the posterior probability $\mathrm{P}(e_1^I | f_1^J)$ is directly modelled as a (log-linear) combination of features, that usually comprise M translational features, and the language model, as in (2):

$$\log \mathrm{P}(e_1^I | f_1^J) = \sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I, s_1^K) + \lambda_{\mathrm{LM}} \log \mathrm{P}(e_1^I) \tag{2}$$

where $s_1^K = s_1, \ldots, s_k$ denotes a segmentation of the source and target sentences respectively into the sequences of phrases $(\hat{f}_1, \ldots, \hat{f}_k)$ and $(\hat{e}_1, \ldots, \hat{e}_k)$ such that (we set $i_0 := 0$):

$$\begin{aligned} \forall k \in [1, K]\ s_k &:= (i_k; b_k, j_k), (b_k \text{ corresponds to starting index of } f_k) \\ \hat{e}_k &:= \hat{e}_{i_{k-1}+1}, \ldots, \hat{e}_{i_k}, \\ \hat{f}_k &:= \hat{f}_{b_k}, \ldots, \hat{f}_{j_k} \end{aligned}$$

Each feature h_m in Eq. 2 can be rewritten as in (3):

$$h_m(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) \tag{3}$$

In theory, log-linear PB-SMT features can apply to the entire sentence, but in practice, those features apply to a single phrase-pair $(\hat{f}_k, \hat{e}_k)$. Thus translational features in (2)

can be rewritten as in (4):

$$\sum_{m=1}^{M} \lambda_m \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) = \sum_{k=1}^{K} \tilde{h}(\hat{f}_k, \hat{e}_k, s_k), \quad \text{with } \tilde{h} = \sum_{m=1}^{M} \lambda_m \hat{h}_m \tag{4}$$

In Eq. 4, $\hat{h}_m$ is a feature defined on phrase-pairs $(\hat{f}_k, \hat{e}_k)$, and λ_m is the feature weight of $\hat{h}_m$. One intuitively natural feature is the phrase translation log-probability ($\hat{h}_m = \log \mathrm{P}(\hat{e}_k|\hat{f}_k)$) where probabilities are estimated using relative frequency count for a phrase pair $(\hat{f}_k, \hat{e}_k)$ independent of any other context information. Other typical features used in PB-SMT (Koehn et al. 2003) are derived from the inverse phrase translation probability ($\log \mathrm{P}(\hat{f}_k|\hat{e}_k)$), the lexical probability ($\log \mathrm{P}_{\mathrm{lex}}(\hat{e}_k|\hat{f}_k)$), and its inverse ($\log \mathrm{P}_{\mathrm{lex}}(\hat{f}_k|\hat{e}_k)$). Our context-informed model will be expressed as additional features in the model.

5 Context-informed features

We conjecture that context-dependent 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. The size of this distribution is possibly limited, and would ideally omit improbable or irrelevant target phrase translations that the standard PB-SMT approach would normally include.

A context-informed feature $\hat{h}_{\mathrm{mbl}}$ can be viewed as the conditional probability of the target phrase $\hat{e}_k$ given the source phrase $\hat{f}_k$ and its context information (CI), as in (5):

$$\hat{h}_{\mathrm{mbl}} = \log \mathrm{P}(\hat{e}_k|\hat{f}_k, \mathrm{CI}(\hat{f}_k)) \tag{5}$$

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

5.1 Lexical features

Lexical features include the immediately neighbouring words within l token positions to the left and right (resp. $f_{i_k-l} \ldots f_{i_k-1}$ and $f_{j_k+1} \ldots f_{j_k+l}$) of a given focus phrase $\hat{f}_k = f_{i_k} \ldots f_{j_k}$. Lexical features thus form a window of size $2l$. The lexical contextual information ($\mathrm{CI}_{\mathrm{lex}}$) can be described as in (6):

$$\mathrm{CI}_{\mathrm{lex}}(\hat{f}_k) = \{f_{i_k-l}, \ldots, f_{i_k-1}, f_{j_k+1}, \ldots, f_{j_k+l}\} \tag{6}$$

Taking the example sentence '*Can you play my favourite old record?*' from Fig. 1, for the single-word focus phrase '*play*', $\mathrm{CI}_{\mathrm{lex}} = \{\mathrm{can}, \mathrm{you}, \mathrm{my}, \mathrm{favourite}\}$ with $l = 2$, i.e. a left and right context of two neighbouring words.

5.2 Part-of-speech tags

In addition to the immediate lexical neighbours of a phrase it is possible to exploit other linguistic information sources characterizing the context. For example, we may consider the part-of-speech (POS) tags of the context words, as well as of the focus phrase itself. In our model, the POS tag of a multi-word focus phrase is the concatenation of the POS tags of the words composing that phrase. We generate a window of size $2l + 1$ features, including the concatenated complex POS tag of the focus phrase. Accordingly, the POS-based contextual information ($\mathrm{CI}_{\mathrm{pos}}$) is described as in (7):

$$\mathrm{CI}_{\mathrm{pos}}(\hat{f}_k) = \{\mathrm{pos}(f_{i_k-l}), \ldots, \mathrm{pos}(f_{i_k-1}), \mathrm{pos}(\hat{f}_k), \mathrm{pos}(f_{j_k+1}), \ldots, \mathrm{pos}(f_{j_k+l})\} \quad (7)$$

For the example sentence (‘*Can you play my favourite old record?*’) (Fig. 1), the POS-based CI for the focus phrase ‘*play*’ is formed as: $\mathrm{CI}_{\mathrm{pos}} = \{\mathrm{pos}(\textit{can}), \mathrm{pos}(\textit{you}), \mathrm{pos}(\textit{play}), \mathrm{pos}(\textit{my}), \mathrm{pos}(\textit{favourite})\} = \{\mathrm{MD}, \mathrm{PRP}, \mathrm{VB}, \mathrm{PRP\$}, \mathrm{JJ}\}$ (with $l = 2$).

5.3 Supertags

Supertags represent complex linguistic categories that express the specific syntactic behaviour of a word in terms of the arguments it takes, and more generally the syntactic environment in which it appears. In our experiments two types of supertags are employed: those from lexicalized tree-adjoining grammar, LTAG (Bangalore and Joshi 1999), and combinatory categorial grammar, CCG (Steedman 2000). Both the LTAG (Chen et al. 2006) and the CCG (Hockenmaier 2003) supertag sets were acquired from the WSJ section of the Penn-II Treebank using hand-built extraction rules. Here we use both the LTAG (Bangalore and Joshi 1999) and CCG (Clark and Curran 2004) 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.

Figures 2 and 3 show the CCG and LTAG supertags respectively for the example sentence ‘*Can you play my favourite old record?*’ Both CCG and LTAG supertags can be combined into parse trees with specific operations (see Figs. 2 and 3). The supertag of a word encodes potentially long-distance syntactic relations with other words in the

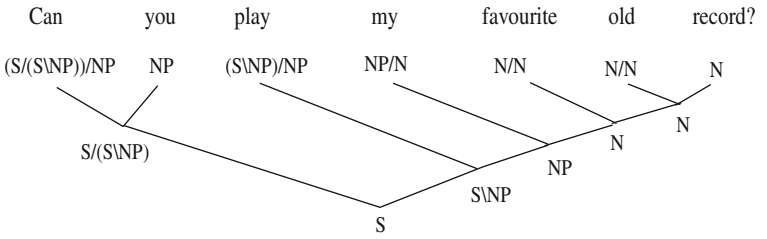

Fig. 2 Example of CCG supertags. CCG supertags are combined under the operations of forward and backward applications into a parse tree

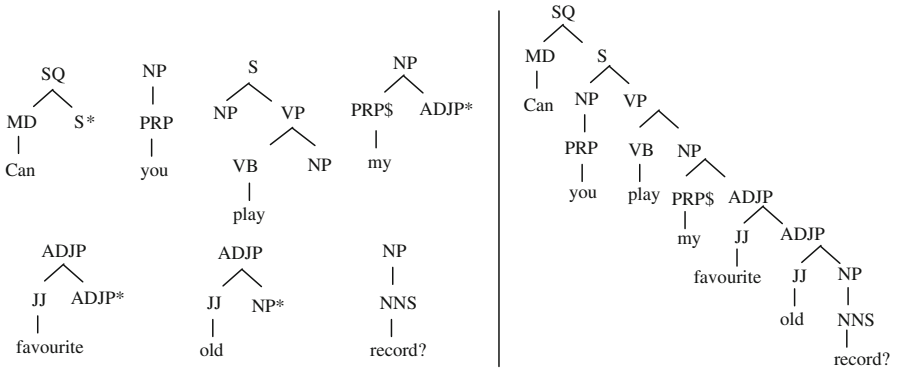

Fig. 3 Example of LTAG supertags. LTAG supertags are combined under the operations of substitution and adjunction into a parse tree

sentence, which could be informative as a feature both of the focus phrase as well as of words neighboring the phrase. As with CI_{pos}, we define the contextual information (CI_{st}) using supertags as in (8):

$$\text{CI}_{\text{st}}(\hat{f}_k) = \{\text{st}(f_{i_k-l}), \ldots, \text{st}(f_{i_k-1}), \text{st}(\hat{f}_k), \text{st}(f_{j_k+1}), \ldots, \text{st}(f_{j_k+l})\} \tag{8}$$

Similar to the CI_{pos} feature, we form the supertag for a multi-word focus phrase by concatenating the supertags of the words composing it. Thus, the supertag-based CI constitutes a window of size $2l + 1$ features. In our experiments, we consider context widths of ± 1 and ± 2 (i.e. $l = 1, 2$) surrounding the focus phrase. For example, the CI of the focus phrase '*play*' in Fig. 2 with CCG supertags is formed as: $\text{CI}_{\text{st}} = \{\text{st}(you), \text{st}(play), \text{st}(my)\} = \{\text{NP}, (\text{S}\backslash\text{NP})/\text{NP}, \text{NP}/\text{N}\}$ (with $l = 1$). We also carried out experiments joining the two supertag types (CCG and LTAG) (cf. Sect. 7.2.2).

5.4 Dependency relations

So far, the context information (CI) of a source phrase ($\hat{f}_k$) is modeled as the sequence of features immediately before and after the focus phrase ($\hat{f}_k$). Although it can be argued that they offer a rich source of information to disambiguate the translation of a source phrase, they remain position-specific and local, and may therefore not provide all information needed for disambiguation. In order to compensate for this, we model position-independent contextual features related to the focus phrase; we choose to model grammatical dependencies linking from and to the head word of the focus phrase ($\hat{f}_k$) with words occurring elsewhere in the sentence.

The identification of the head word of a phrase is non-trivial, as SMT phrases are not restricted to linguistically coherent phrases, so the identification of a head word cannot be done with linguistic rules of thumb (e.g. select the head noun from the noun phrase). In our work, we identify head words of SMT phrases with the use of a dependency tree generated for the sentence. For all words in a given source phrase, the word that occupies hierarchically the highest position in the dependency tree is

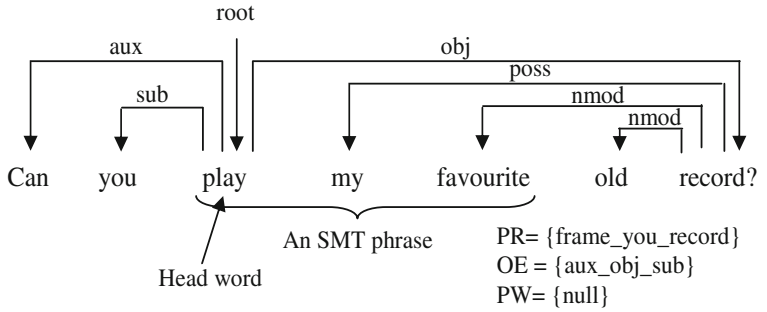

Fig. 4 The dependency parse tree of the English sentence *Can you play my favourite old record?* and the dependency features extracted from it for the SMT phrase *play my favourite*

chosen as the head word. For example, Fig. 4 shows the above English sentence (‘*Can you play my favourite old record?*’) and its dependency parse tree. In Fig. 4 we see that the head word of an English phrase ‘*play my favourite*’ identified by the PB-SMT system is ‘*play*’ according to the tree structure.

We consider the following dependency features, drawing on the syntactic dependencies emanating from or pointing to the head word of the source focus phrase (see also Fig. 4):

OE (outgoing edges)—For the head word of the focus phrase, we extract a list of zero or more relations with other words of which the word is the parent (i.e. the dependency type labels on all modifying dependency relations). The list of relations is concatenated and sorted uniquely and alphabetically into a single feature. This feature is denoted as OE, for ‘outgoing edges’. For example, the head word (‘*play*’) of the focus phrase (‘*play my favourite*’) has three outgoing edges: auxiliary, subject and object (see Fig. 4). Therefore, the OE feature is formed as: OE = {aux_obj_sub}.

PR (parent relation)—For the head word of the focus phrase we extract the relation it has with its parent. If the head word is a verb, then the subcategorization frame information is extracted and used as this feature. This feature is denoted as PR, for ‘parent relation’. For example, the head word *‘play’* is a verb, and so we extract its subcategorization frame information from the tree, namely {frame_you_record}.

PW (parent word)—Extending the PR feature, we encode the identity of the parent word of the head word of the focus phrase. This feature is denoted as PW, for ‘parent word’. For example, the head word *‘play’* is root according to the tree structure, and so we set the PW feature to *null*.

Together we refer to these dependency features as the grammatical *dependency information* ($\mathrm{CI}_{\mathrm{di}}(\hat{f}_k)$) of the focus phrase ($\hat{f}_k$). These dependency features can be applied both individually and jointly. For instance, a combination of three dependency features (PR, OE and PW) defines the contextual information $\mathrm{CI}_{\mathrm{di}}(\hat{f}_k)$ as in (9):

$$\mathrm{CI}_{\mathrm{di}}(\hat{f}_k) = \{\mathrm{OE}, \mathrm{PR}, \mathrm{PW}\} \tag{9}$$

Two published studies are closely related to our work on integrating dependency features. Carpuat and Wu (2007) mention in passing that their WSD system uses basic dependency relations, but the nature of this information is not further described, nor is its effect. Max et al. (2008) exploit grammatical dependency information, in addition to information extracted from the immediate context of a source phrase. Our approach differs with (Max et al. 2008) at least in three respects:

1. Max et al. (2008) select a set of the 16 most informative dependency relations for their experiments. Dependencies are considered that link any of the tokens in the given source phrase to tokens outside the phrase. Each dependency type is represented in the vector by the outside word it involves, or by the symbol 'nil', which indicates that this type of dependency does not occur in the phrase under consideration. In contrast to this approach, we used all (26) dependency relations in our experiments, while only extracting features from the head words of the SMT phrases.
2. They filter out phrases from the phrase table for which $P(\hat{e}_k|\hat{f}_k) < 0.0002$. In contrast, we keep all phrase pairs.
3. Their experimental data contains 95K English-to-French training sentence pairs, while we carried out a range of experiments considering different data sizes, domains, and language pairs, elaborated further in Sect. 7.

We compare the dependency features with incorporating words, part-of-speech tags and supertags as context, in order to observe the relative effects of position-independent and position-dependent features. While supertags represent an abstract view upwards the tree or graph, excluding other lexical nodes and anything below the lowest common ancestors in the tree between other lexical nodes and the words captured in the contextual features, dependency relations directly encode relations between tokens. One can follow a dependency and retrieve the lexical modifier or head at any distance.

5.5 Semantic roles

Semantic role labeling is an established benchmark task in NLP research since the 2004 CoNLL shared task (Carreras and Márquez 2004). The task is to identify the semantic arguments associated with a clause's predicate, and their classification into specific semantic roles with respect to the predicate. Typical roles include agent, theme, temporal or locative modifier, etc. The CoNLL 2008 shared task (Surdeanu et al. 2008) introduced a unified dependency-based formalism that modeled both syntactic dependencies and semantic roles for English. A dependency-based semantic role labeler (SRL) (Johansson and Nugues 2008) finds all semantic graphs around each predicate verb or noun in an input sentence in addition to the dependency parse tree.

Recently, Wu and Fung (2009) utilized semantic roles in improving SMT accuracy by enforcing consistency between the semantic predicates and arguments across both the input sentence and the translation output. Inspired by Wu and Fung (2009), we introduce semantic information as a new contextual feature in PB-SMT for a English-to-Dutch translation task. In order to obtain the semantic graph information of English

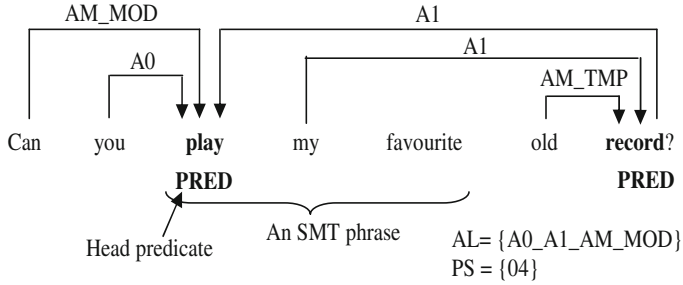

Fig. 5 The semantic graph of an English sentence and the semantic features extracted from it for an SMT phrase

sentences, we used the LTH semantic parser[1] (Johansson and Nugues 2008), which assigns both predicative (PropBank-based) and nominative (NomBank-based) graphs.

The semantic information (CI_{si}) of a source phrase ($\hat{f}_k$) originates from the semantic (verbal or nominal) predicate captured in that phrase. This introduces three possible cases: (a) there is no predicate in the source phrase, (b) there is only one predicate in the source phrase, in which case this is chosen as the head predicate to define (CI_{si}), and (c) more than one predicate occurs in the source phrase; for such cases, the predicate that occupies the hierarchically superior position in the dependency parse tree is chosen as the head predicate. For example, Fig. 5 shows a English sentence ('*Can you play my favourite old record?*') and two semantic graphs around its two predicates (verbal: '*play*' and nominal: '*record*') identified by SRL. Figure 5 also shows an English phrase '*play my favourite*' identified by our baseline PB-SMT system, Moses, which contains only the verbal predicate '*play*'.

In our experiments, two semantic features were generated for the head predicate:

AL (argument labels)—We extract the list of one or more predicate roles (argument labels) of the head predicate of a source phrase. The list of roles is concatenated and sorted uniquely and alphabetically into a single feature. This feature is denoted as AL, for 'argument labels'. For example, Fig. 4 illustrates that the head predicate ('*play*') of the focus phrase ('*play my favourite*') has three semantic dependencies (argument labels): acceptor (A0), thing accepted (A1) and modal (AM_MOD).

PS (predicate sense)—In addition to the semantic roles of a predicate, SRL attempts to disambiguate the sense of the predicate in the source sentence. We extract the sense of the head predicate of the source phrase. This feature is denoted as PS, for 'predicate sense'. For example, the sense of the head predicate ('*play*') in the sentence in Fig. 5 is {04}.

The two features AL and PS are applied both individually and jointly. For instance, a combination of two semantic features defines the contextual information $CI_{si}(\hat{f}_k)$ of the source phrase ($\hat{f}_k$) as in (10):

$$CI_{si}(\hat{f}_k) = \{AL, PS\} \tag{10}$$

[1] http://nlp.cs.lth.se/software/.

6 Memory-based classification

Stroppa et al. (2007) pointed out that directly estimating context-dependent phrase translation probabilities using relative frequencies is problematic. Indeed, Zens and Ney (2004) showed that the estimation of $P(\hat{e}_k|\hat{f}_k)$ using relative frequencies results in overestimation of the probabilities of long phrases; consequently, 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, this estimation problem can only become worse.

As an alternative, we make use of memory-based machine learning classifiers that are able to estimate $P(\hat{e}_k|\hat{f}_k, CI(\hat{f}_k))$ by similarity-based reasoning over memorized nearest-neighbour examples of source–target phrase translations to a new source phrase to be translated. In this work, we use two approximate memory-based classifiers: IGTree and TRIBL[2] (Daelemans and van den Bosch 2005). Both algorithms approximate the unabridged (and computationally expensive) memory-based or k-nearest neighbour (k-NN) classifier (Aha et al. 1991).

6.1 Fast approximate memory-based classification: IGTree

IGTree makes a heuristic approximation of k-NN search (Aha et al. 1991) 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. 1997a). As a normal k-NN classifier, IGTree retains the labeling information of all training examples, but in a compressed form. 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 concatenated single values, just as its POS tags or supertags are) and its contextual information, associated with a symbolic class label representing the associated target phrase found through an alignment procedure. A weighting metric such as information gain (IG) is used to determine the order in which features are tested in the tree (Daelemans et al. 1997a). The source phrase itself is intuitively the feature with the highest prediction power; it should take precedence in the similarity-based reasoning, and indeed it does, as it always receives the highest IG value.

Prediction in IGTree is a straightforward traversal of the decision tree from the root node down, where a step is triggered by a match between a feature value of the new example and an arc fanning out of the current node. When traversal 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, this would be 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. As the source

[2] An implementation of IGTree and TRIBL is freely available as part of the TiMBL software package, which can be downloaded from http://ilk.uvt.nl/timbl.

phrase in focus is the first feature tested, when an input mismatches on the source phrase, the prior target phrase distribution in the training set is returned.

6.2 A hybrid between k-NN and IGTree: TRIBL

TRIBL, which stands for Tree-based approximation of Instance-Based Learning, a hybrid combination of IGTree and unabridged k-NN classification, performs heuristic approximate nearest neighbour search (Daelemans et al. 1997b). A parameter n determines the switching point in the feature ordering from IGTree to normal k-NN classification. The TRIBL approximation performs an initial decision-tree split of the database of training examples on the n most informative features, like IGTree would. Throughout our experiments we set $n = 1$; thus, we split on the values of the single most informative feature according to information gain (IG). During classification, after sub-selecting training examples matching on the most informative feature, the nearest-neighbour distance function is applied to the remaining features (weighted by their IG) to arrive at the set of nearest-neighbours. When predicting a target phrase given a source phrase and its context information, the identity of the source phrase is effectively (and also intuitively) the feature with the highest prediction power. This implies that nearest neighbours always match on the source phrase, and are most similar (preferably, identical) with respect to their contextual features.

A parameter k determines the k closest radii of distances around the source phrase that encompass the nearest neighbours; then, the distribution of target phrases associated with these nearest neighbours is taken as the output of the classification step. The contribution of a single nearest neighbour in this set can be weighted by its distance to the source phrase to be translated, e.g. by assigning higher weights to closer neighbours. In our experiments, we empirically set the value of k, and use exponential decay for the distance-weighted class voting (Daelemans and van den Bosch 2005). Choosing the optimal setting of k can be handled empirically, as are other hyperparameter settings for the k-NN part of TRIBL. We used a heuristic-automated hyperparameter estimation method based on wrapped progressive sampling (Van den Bosch 2004)[3] throughout our experiments. Thus, TRIBL's hyperparameters are retuned with each experiment.

A TRIBL classification produces a class distribution derived from the aggregate distance-weighted class voting generated by all found nearest neighbours, from which we estimate $P(\hat{e}_k | \hat{f}_k, CI(\hat{f}_k))$ for all possible $\hat{e}_k$. By normalizing the class votes generated by TRIBL, we obtain the posterior probability distributions we are interested in.

Although TRIBL is a fast approximation of unrestricted k-NN, it retains its relatively large memory consumption, which becomes hard to handle on current computing machinery when the number of examples is of the order of 10^7 or higher. In the experiments reported in this paper, experiments with small-scale data sets are typically performed with TRIBL; experiments on large-scale data sets are performed

[3] http://ilk.uvt.nl/paramsearch/.

with IGTree. In Sect. 7.4.1 we compare the two approaches directly in a learning curve study.

6.3 Feature integration

The output of memory-based k-NN classification is a set of weighted class labels, representing the possible target phrases ($\hat{e}_k$) given a source phrase ($\hat{f}_k$) and its context information (CI). Once normalized, these weights can be seen as the posterior probabilities of the target phrases ($\hat{e}_k$) which give access to $\mathrm{P}(\hat{e}_k|\hat{f}_k, \mathrm{CI}(\hat{f}_k))$. Thus, from the classifier's output we can derive the feature $\hat{h}_{\mathrm{mbl}}$ defined in Eq. 5. In addition to $\hat{h}_{\mathrm{mbl}}$, we derive a simple two-valued feature $\hat{h}_{\mathrm{best}}$, defined as in Eq. 11:

$$\hat{h}_{\mathrm{best}} = \begin{cases} 1 & \text{if } \hat{e}_k \text{ maximizes } \mathrm{P}(\hat{e}_k|\hat{f}_k, \mathrm{CI}(\hat{f}_k)) \\ \approx 0 & \text{otherwise} \end{cases} \tag{11}$$

where $\hat{h}_{\mathrm{best}}$ is set to 1 when $\hat{e}_k$ is one of the target phrases with highest probability according to $\mathrm{P}(\hat{e}_k|\hat{f}_k, \mathrm{CI}(\hat{e}_k))$; otherwise, $\hat{h}_{\mathrm{best}}$ is set to 0.000001. We performed experiments by integrating these two features $\hat{h}_{\mathrm{mbl}}$ and $\hat{h}_{\mathrm{best}}$ directly into the log-linear model of Moses. Their weights are optimized using minimum error-rate training (MERT)(Och 2003) on a held-out development set for each of the experiments.

As Stroppa et al. (2007) point out, PB-SMT decoders such as Pharaoh (Koehn 2004a) or Moses (Koehn et al. 2007) rely on a static phrase 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 phrases occur, no context information is kept in the phrase table, and there is no way to recover this information from the phrase table.

In order to take into account the context-informed features within such decoders, the test set to be translated is preprocessed. Each word appearing in the test set (or, during development, in the development set) is assigned a unique identifier. First we derive the phrase table from the training data. Subsequently, we generate all possible phrases from the test set. These phrases are then looked up in the phrase table, and when found, the phrase along with its contextual information is given to the memory-based classifier to be classified. As stated above, memory-based classifiers produce target phrase distributions according to the training examples found within the k-nearest distance radii around the source phrase to be classified. We derive target phrase probabilities from this distribution and temporarily insert them into a new phrase table with the original phrase table estimates, to directly take our feature functions into account in the log-linear model. Thus we create an updated phrase table.

A lexicalized reordering model is used for all the experiments undertaken on the development and test sets. The source phrases in the reordering table are replaced by the sequence of unique identifiers when the new phrase table is created. After replacing all words by their unique identifiers, we perform MERT using our updated phrase table to optimize the feature weights.

6.4 Classification example

In this section, we give an example to illustrate how a source phrase with given additional context information is classified into a distribution over possible target phrases. We consider a particular contextual feature (CCG±1) to illustrate the classification task. We select a source English sentence (‘*let me make a suggestion to the commission as to how this problem could be tackled*’) from the development set of the English-to-Spanish translation task.[4] This sentence contains the ambiguous single-word phrase ‘*make*’, of which the contextually appropriate Spanish translation would be ‘*hacer*’.

Following Eq. 8, we take the CCG supertags of the neighbouring (±1) words around the source phrase ‘*make*’ in order to form its context information, namely: CI (*make*) = {st(*me*), st(*make*), st(*a*)} = {(S\NP)/NP, NP, NP/N}. Thus, we form a test example $\langle make, \text{CI} \rangle$ which is given to the classifier for classification. As part of the earlier offline training phase, millions of training examples are generated that take the same form as the test examples, labeled with classes (aligned target phrases); one example for each alignment in the training data. A memory-based classifier is then trained on these millions of training examples.

During decoding, we generate all possible test examples from the test set and give them to the classifier in order to obtain possible translations of the source phrases. For the above test example ($\langle make, \text{CI} \rangle$) we classify it with the relevant classifier, which gives us a class distribution in the form of a list of target phrases which are context-sensitive translations of the source phrase ‘*make*’. A weight is associated with each class. From these weights we estimate the probabilities of translation into target phrases ($\hat{e}_k$) from the source phrase ‘*make*’ with its additional context information, which are $\text{P}(\hat{e}_k|\text{make, CI(make)})$ (where k denotes the distribution size). According to the classification result, we found that ‘*hacer*’ is the translation with the highest weight for the source phrase ‘*make*’, i.e. the translation probability (P(hacer|make, CI(make))) is the highest. Table 7 shows some of the possible Spanish translations of the English phrase ‘*make*’ including ‘*hacer*’ with their memory-based context-dependent translation probabilities (i.e. memory-based scores: $\text{P}(\hat{e}_k|\hat{f}_k, \text{CI}(\hat{f}_k))$) compared with context-independent translation probabilities (i.e. baseline scores: $\text{P}(\hat{e}_k|\hat{f}_k)$).

Table 7 shows that the memory-based classifier assigns a relatively higher score to the most suitable Spanish phrase ‘*hacer*’, while it assigns lower scores to three of the four alternative translations listed. The baseline phrase translation probability is estimated using the relative frequency counts of source and target phrases. Additionally we report the target phrase distribution size (TPDS, bottom line in Table 7) for the source phrase ‘*make*’ in the baseline system, 388 phrases, as well as in our memory-based model, 181 phrases. This illustrates how the memory-based classifier typically produces a reduced set of target phrases for a given source phrase in context.

As an additional point of analysis, we also compared the log-linear weights (λ_i) of the context-informed memory-based features $\hat{h}_{\text{mbl}}$ (cf. Eq. 5) and $\hat{h}_{\text{best}}$ (cf. Eq. 11) with the weight of a baseline feature derived from the forward phrase translation probability ($\hat{h}_{\text{base}} = \log \text{P}(\hat{e}_k|\hat{f}_k)$) for the above experiment (CCG±1, an English-to-Spanish

[4] The English-to-Spanish translation task is reported in Sect. 7.4.1.

Table 7 Some of the possible Spanish translations of the English phrase *make* with their memory-based context-dependent translation probabilities (rightmost column) compared against context-independent translation probabilities of the baseline system

	Baseline $P(\hat{e}_k \mid \hat{f}_k)$	CCG±1 $P(\hat{e}_k \mid \hat{f}_k, CI\hat{f}_k)$
hacer	0.2353	0.3412
hagan	0.0851	0.0057
realizar	0.0277	0.0305
hacen	0.0245	0.0073
haga	0.0142	0.0113
...	...	...
TPDS	388	181

TPDS target phrase distribution size

Table 8 Weights of different log-linear features of the CCG±1 system

$\hat{h}_{\text{base}}$	$\hat{h}_{\text{mbl}}$	$\hat{h}_{\text{best}}$
0.0263951	0.0463334	0.00507119

translation task described further in Sect. 7.4.1). The λ_i of the various log-linear features directly affects the phrase scores during translation; thus, λ_i plays a crucial role in selecting the most appropriate candidate phrases.

Table 8 indicates that our context-informed models ($\hat{h}_{\text{mbl}}$, $\hat{h}_{\text{best}}$) contribute positively to the phrase-scoring process during translation. Moreover, MERT (Och 2003) assigns a notably higher weight to the context-informed feature $\hat{h}_{\text{mbl}}$ than the baseline feature $\hat{h}_{\text{base}}$, directly indicating the importance of the memory-based context-informed models.

7 Experiments and results

We carried out experiments by systematically applying various source-side contextual features for different language pairs with varying training data sizes. The system outputs are evaluated across a wide range of automatic evaluation metrics: BLEU (Papineni et al. 2002), NIST (Doddington 2002), METEOR (Lavie and Agarwal 2007), TER (Snover et al. 2006), WER, and PER. Additionally we performed statistical significance tests using bootstrap resampling methods on BLEU (Koehn 2004b). The confidence level (%) of the improvements obtained by the best performing context-informed systems with respect to the PB-SMT baseline are reported. An improvement in system performance at a confidence level above 95% is assumed to be statistically significant.

We divide the reports on our experiments into five subsections. Section 7.1 provides an overview of statistical properties of the corpora we used in our experiments. Section 7.2 reports on small-scale data sets representing the language pairs Dutch-

to-English, English-to-Hindi, and English-to-Czech, with less than 300,000 training examples. Section 7.3 reports on large-scale data sets with more than 500,000 training sentence-pairs, representing the language pairs Dutch-to-English, English-to-Dutch, English-to-Japanese, and English-to-Chinese. In Sect. 7.4, we present the results of learning curve experiments we carried out on three different language pairs: English-to-Spanish, Dutch-to-English, and English-to-Dutch. In Sect. 7.5 we present a manual qualitative analysis of the MT outputs.

7.1 Data

The various corpora we used for carrying out our experiments are listed in Table 9. In this table we list for each data set the number of sentences, source (S) and target (T) vocabulary size (VS), and average sentence length (ASL). The corpora have the following origins:

Table 9 Corpora statistics

Data source	Data set	Sentences	VS (S)	VS (T)	ASL (S)	ASL (T)
Dutch-to-English: open subtitles	Train.	286,160	59,863	44,594	7.34	8.85
	Dev.	1,000	1,286	1,228	7.00	8.02
	Eval.	1,000	1,435	1,327	6.67	7.29
English-to-Hindi: EILMT	Train.	6,755	16,344	24,734	24.56	25.97
	Dev.	500	3,689	4.021	24.58	25.99
	Eval.	495	3,598	3,879	24.07	25.51
English-to-Czech: WMT 2010	Train.	94,501	89,672	241,948	20.91	18.68
	Dev.	2,051	10,531	14,757	21.18	17.74
	Eval. 1	2,525	11,945	17,817	21.64	17.73
	Eval. 2	2,489	12,326	13,726	21.70	18.57
Dutch-to-English: Europarl	Train.	1,311,111	247,079	126,141	26.97	27.74
	Dev.	1,000	4,020	3,450	24.81	25.17
	Eval.	1,000	4,312	3,751	24.95	26.15
English-to-Japanese: NTCIR-8	Train.	600,000	193,526	65,934	28.30	29.58
	Dev.	1,814	7,722	4,911	29.36	37.09
	Eval. 1	927	4,264	3,574	28.77	37.08
	Eval. 2	1,119	5,157	3,062	28.51	35.57
English-to-Chinese: NIST-08	Train.	500,000	112,773	128,647	34.38	31.99
	Dev.	1,082	5,097	5,693	33.29	28.12
	Eval.	1,357	2,990	3,320	29.38	24.48
English-to-Spanish: Europarl	Train.	1,639,764	260,099	980,000	24.27	25.96
	Dev.	2,000	8,651	10,371	26.54	27.28
	Eval.	2,000	8,689	10,552	27.15	27.90

S source, *T* target, *VS* vocabulary size, *ASL* average sentence length

Dutch-to-English (Open Subtitles): This corpus is collected as part of the Opus collection of freely available parallel corpora (Tiedemann and Nygaard 2004).[5] The corpus contains user-contributed translations of movie subtitles.

English-to-Hindi (EILMT): This small EILMT tourism domain corpus was released for the shared task on English-to-Hindi SMT (Venkatapathy 2008).[6]

English-to-Czech (WMT 2010): For the English-to-Czech translation task we employed the News Commentary training data set released in the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT-MetricsMATR 2010)[7]. To tune the system during development, we used the WMT 2008 test set of 2,051 sentences. For evaluation purposes we used two different test sets: the WMT 2009 test set of 2,525 sentences, and the WMT 2010 testset of 2,489 sentences.

Dutch-to-English (Europarl): The Dutch-to-English Europarl parallel corpus is extracted from the proceedings of the European Parliament (Koehn 2005).[8]

English-to-Japanese (NTCIR-8): The experimental data sets were taken from the NTCIR-8 Patent Translation Task.[9] For the purpose of evaluation, we used two different test sets: the first test set contains 927 sentences (henceforth referred to as 'EJTestset1'), and the second test set contains 1,119 sentences ('EJTestset2').

English-to-Chinese (NIST-08): Our English-to-Chinese training text contains sentence pairs of benchmark news text from the NIST Open Machine Translation 2009 Evaluation (MT09).[10] We used the NIST MT05 test set sentences (development set) for tuning, and the NIST MT08 'current' test set sentences for evaluation.

English-to-Spanish (Europarl): This training corpus was provided in the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT-MetricsMATR 2010).[11] We use the WMT 'test2006' set as a development set, and the WMT 'test2008' set as the test set; both sets contain 2,000 sentence pairs.

7.2 Experiments on small-scale data sets

7.2.1 Dutch-to-English

Small-scale experiments were performed on the Dutch-to-English Open Subtitles corpus (cf. Sect. 7.1). To generate dependency features, the Dutch sentences were parsed using Frog,[12] a robust morphosyntactic analyzer and dependency parser (Van den Bosch et al. 2007) that generates approximately the double amount of errors in labeled relations as compared to the English equivalent dependency parser used in our study.

[5] http://urd.let.rug.nl/tiedeman/OPUS/OpenSubtitles.php.

[6] http://ltrc.iiit.ac.in/nlptools2008/index.html.

[7] http://www.statmt.org/wmt10/.

[8] http://www.statmt.org/europarl/.

[9] http://www.cl.cs.titech.ac.jp/~fujii/ntc8patmt/.

[10] http://www.itl.nist.gov/iad/mig//tests/mt/2009/.

[11] http://www.statmt.org/wmt10/.

[12] http://ilk.uvt.nl/frog/.

Table 10 Experiments with words and parts-of-speech as contextual features

Experiments	BLEU	NIST	METEOR	TER	WER	PER
Baseline	*32.39*	*6.11*	*55.39*	*50.15*	*49.67*	*43.12*
Word±2	32.48	6.11	55.72	50.40	50.43	42.91
POS±2	33.07	6.13	56.17	50.07	49.38	42.85
POS±2†	33.29 (74%)	6.17	55.72	49.56	48.91	42.77
Word±2 + POS±2	32.59	6.09	55.36	50.11	49.63	43.10

Table 11 Experiments with dependency relations

Experiments	BLEU	NIST	METEOR	TER	WER	PER
Baseline	*32.39*	*6.11*	*55.39*	*50.15*	*49.67*	*43.12*
PR	32.69	6.08	55.08	50.48	50.11	43.58
OE	32.61	6.00	55.53	52.40	51.56	45.09
PR+PW	32.74	6.06	55.98	51.15	50.75	43.61
PR+OE	33.06 (60%)	6.20	55.70	49.45	48.83	42.44
PR+OE+PW	32.79	6.18	55.37	49.51	49.03	42.43

We performed three series of experiments using the TRIBL classifier. In the first series, words, parts-of-speech, and their combination are added as contextual information. The experimental results are reported in Table 10. In all cases, the width of the left and right contexts is 2. An additional experiment (labeled POS±2†)[13] was performed in which the concatenated parts-of-speech of the focus phrases were not included as a feature. As can be observed from Table 10, the POS±2† experiment produces the best improvements (0.90 BLEU points; 2.78% relative) over the baseline. However, the improvement is not statistically significant.

A second series of experiments was performed involving dependency relations as source context. The results are shown in Table 11. Five experiments were performed combining dependency features (outgoing edges: OE, parent relation: PR, and parent word: PW). The combination of PR and OE produces the best results in terms of BLEU and NIST: we observe a 0.67 absolute improvement corresponding to a 2.07% relative improvement in terms of BLEU, which is not statistically significant.

In the third series we combined the position-independent PR+OE dependency features with the position-dependent word and part-of-speech features. The combined experimental results are reported in Table 12. We observe that combining POS±2† with PR+OE yields the highest BLEU improvement (1.0 BLEU point; 3.08% relative) over the baseline, which is statistically significant at a 98.7% level of confidence. The best METEOR score (an improvement of 1.18 METEOR points over the baseline; 2.14% relative) is obtained when PR+OE is combined with POS±2.

[13] Signalling one particular exception, we use the *dag*(†) symbol for experiments in which syntactic information of the focus phrase is ignored.

Table 12 Experiments combining dependency relations, words and part-of-speech

Experiments	BLEU	NIST	METEOR	TER	WER	PER
Baseline	*32.39*	*6.11*	*55.39*	*50.15*	*49.67*	*43.12*
PR+OE+Word±2	33.05	6.11	56.02	50.62	49.82	43.68
PR+OE+POS±2	33.30	6.09	56.57	50.52	50.17	43.81
PR+OE+POS±2†	33.39 (98.7%)	6.11	56.30	50.43	50.34	43.54

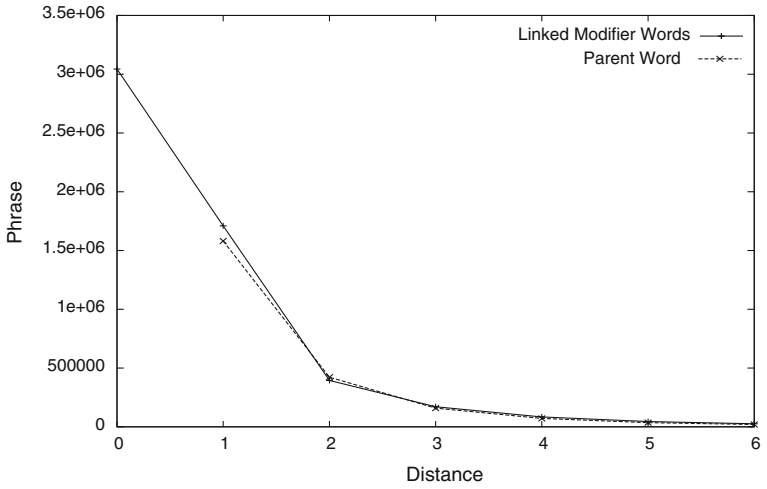

Fig. 6 Distances found between phrase boundaries with linked modifier words and with parent words

In sum, the small-scale Dutch-to-English translation task shows the POS contextual feature to produce the largest single-feature improvement over the baseline, while the difference in score between POS-based and dependency-based contextual models is negligible. Moreover, the highest improvement in BLEU over the baseline is obtained employing the combination of the POS- and dependency-based features, which is statistically significant.

As an additional analysis, Fig. 6 displays the distribution of distances (number of tokens) between the source phrase boundary and the words outside the phrase linked through a dependency relation. There are about twice as many outgoing modifier dependency relations linking to modifier words outside the focus phrase than to phrase-internal modifiers. About half of the phrases have the root of the dependency graph as the parent, i.e. they are the main verbs. For the remaining phrases, the parent of the head-word is a phrase-external word. From the distance distribution statistics, we find that the average distance of head-modifying words to the phrase boundary is only 0.75 tokens when including phrase-internal relations, indicating that modifiers of the phrase are usually not too far away, and are mostly immediate neighbours. In contrast, parent words of the phrase's head word are found relatively further away, at an average distance of 1.69 tokens outside the phrase boundary.

Table 13 Experiments applying individual features in English-to-Hindi translation

Experiments	BLEU	NIST	METEOR	TER	WER	PER
Baseline	*10.93*	*4.54*	*28.59*	*74.87*	*82.06*	*56.67*
Word±1	10.76	4.53	28.27	75.42	82.75	57.26
Word±2	11.24 (85%)	4.58	28.27	74.89	82.18	56.45
POS±1	10.82	4.55	28.59	74.76	81.85	56.65
POS±2	11.00	4.55	28.90	74.67	81.89	56.72
CCG±1	11.14	4.58	27.94	74.84	82.19	56.76
CCG±2	11.07	4.57	28.59	74.76	81.85	56.65
LTAG±1	11.19	4.55	28.28	74.67	81.48	56.78
LTAG±2	11.17	4.57	28.59	74.73	81.98	56.63
CCG-LTAG±1	11.01	4.53	28.73	75.34	82.62	56.89
CCG±1+LTAG±1†	11.04	4.55	28.73	74.94	82.14	56.66
Super-pair±1†	11.02	4.58	27.62	74.45	81.45	56.45
Super-pair±2†	11.15	4.58	28.27	74.87	82.22	56.45
PR+OE	11.02	4.58	28.28	74.65	81.75	56.55

7.2.2 English-to-Hindi

For the English-to-Hindi translation pair we conduct experiments using supertags and dependency context features, both individually and jointly, comparing these to using lexical and POS contextual features. Like other Indian languages, Hindi is a free word order language. Experiments were carried out using the relatively small EILMT tourism corpus (cf. Sect. 7.1). In order to obtain the dependency parse information for English sentences, the Malt dependency parser[14] (Nivre et al. 2006) was used. The experimental results for individual and joint features, using the TRIBL classifier, are displayed in Tables 13 and 14, respectively.

For the English-to-Hindi translation task, we copied the previously best-performing set-up obtained from the Dutch-to-English translation task. Experimental results in Table 13 show that among the homogeneous features, Word±2 produces the best improvement (0.31 BLEU points, 2.84% relative) over the baseline. This improvement is not statistically significant, yet close to the significance level. Other context-informed features also produce small but consistent improvements over the baseline in terms of BLEU. However, these improvements with respect to the baseline are not statistically significant. The other evaluation metrics show similar improvements.

The results of combining the dependency features PR+OE with various contextual features are shown in Table 14. Combining PR+OE with POS±2 and the concatenation of CCG and LTAG supertag features, referred to as PR + OE + POS ± 2 + CCG + LTAG ± 1† (see last row in Table 14), we achieve the overall best improvement (0.41 BLEU points; 3.7% relative) over the baseline. Again, the improvement is not statisti-

[14] http://maltparser.org/download.html.

Table 14 Experiments applying combinations of features in English-to-Hindi translation

Experiments	BLEU	NIST	METEOR	TER	WER	PER
Baseline	*10.93*	*4.54*	*28.59*	*74.87*	*82.06*	*56.67*
PR+OE+$\cdots$						
POS±2	11.08	4.56	28.59	75.39	82.67	56.65
Word±2	11.02	4.55	28.59	75.13	82.27	56.85
CCG±1	11.02	4.57	28.27	74.77	82.08	56.56
LTAG±1	11.08	4.56	28.27	75.39	82.67	56.65
CCG±1+LTAG±1†	11.02	4.57	28.27	74.77	82.08	56.56
Super-pair±1†	11.02	4.54	28.27	75.00	82.00	56.82
Super-pair±2†	11.23	4.57	28.59	74.74	81.92	56.55
POS±2+CCG+LTAG±1†	11.34 (89%)	4.58	27.94	74.70	82.00	56.44

cally significant, yet is close to the significance level. Similar trends are observed on other evaluation metrics for the combined features.

In sum, the word contextual model produces the biggest single-feature improvement over the baseline in terms of BLEU. The combination of dependency, supertag, and POS features brings about the highest BLEU score in this translation task, although the improvements over the baseline are not statistically significant.

7.2.3 English-to-Czech

For the English-to-Czech translation task (Penkale et al. 2010) we employed the News Commentary training data set (cf. Sect. 7.1). We employed the previously best performing set-up in terms of context width and feature combinations, and the TRIBL classifier. The evaluation results on the WMT 2009 test set are reported in Table 15. We observe that small improvements over the Moses baseline are achieved for CCG±1, PR and PR+OE features in terms of BLEU. Moderate improvements in the METEOR and TER evaluation scores are achieved for all features except LTAG±1 and Word±2. The highest METEOR score over the baseline is obtained for the dependency parent relation (PR: 0.31 METEOR points improvement; 0.91% relative). On the TER evaluation metric, the best performing set-up, CCG±1, yields an absolute reduction of 0.42 TER points over the baseline. Similar trends are also observed for WER and PER distance metrics. Moreover, gains for the CCG±1 and PR+OE features over the baseline are seen across the most evaluation metrics.

Experimental results on the WMT 2010 test set are shown in Table 16. We observe that the improvements on this test set are similar to the improvements on the WMT 2009 test set. CCG±1 yields the highest improvements across all evaluation metrics except METEOR. As far as the METEOR evaluation metric is concerned, the Super-Pair±1 feature produces the best improvement (0.32 METEOR points gains, 0.92% relative) over the baseline. CCG±1 yields a 0.21 BLEU points gain (2.68% relative increase) and a 0.43 TER points reduction over the baseline. In contrast, POS±2 and Word±2 do not bring about any improvements across any of the evaluation metrics.

Table 15 Experimental results on the WMT 2009 test set

Experiments	BLEU	NIST	METEOR	TER	WER	PER
Baseline	*7.83*	*3.90*	*34.13*	*87.66*	*80.53*	*67.88*
CCG±1	7.88	3.95	34.23	87.24	80.15	67.39
LTAG±1	7.67	3.89	34.00	87.90	80.86	68.00
CCG-LTAG±1	7.80	3.90	34.35	88.24	81.17	68.16
Super-pair±1	7.82	3.90	34.38	87.96	80.84	68.18
POS±2	7.80	3.90	34.25	87.87	80.84	67.95
Word±2	7.50	3.83	33.84	88.73	81.68	68.67
PR	7.85	3.92	34.44	87.53	80.57	67.88
PR+OE	7.86	3.92	34.29	87.55	80.53	67.80

Table 16 Experimental results on the WMT 2010 test set

Experiments	BLEU	NIST	METEOR	TER	WER	PER
Baseline	*8.05*	*3.97*	*34.61*	*86.01*	*78.54*	*67.48*
CCG±1	8.26	4.02	34.76	85.58	78.06	66.96
LTAG±1	8.00	3.95	34.57	86.41	78.95	67.72
CCG-LTAG±1	8.09	3.96	34.90	86.62	79.18	67.91
Super-Pair±1	8.11	3.95	34.93	86.62	79.05	68.08
POS±2	7.91	3.94	34.57	86.5	79.03	67.84
Word±2	7.57	3.88	34.16	87.13	79.77	68.39
PR	8.06	4.00	34.89	85.98	78.62	67.43
PR+OE	8.03	3.99	34.81	85.97	78.63	67.44

In sum, slight improvements over the baseline are seen for supertag and dependency features, while POS and word contexts do not improve the baseline at all. None of the improvements over the baseline models are statistically significant in terms of BLEU.

If we compare the effectiveness of the various contextual features both collectively and individually, in all small-scale translation tasks we see that supertags seem to be the most effective context features, as compared to neighbouring words and part-of-speech. Arguably, one can surmise that the differences in word order between the source and target languages in our experiments are best treated by a feature which is not entirely restricted to local context.

7.3 Experiments on large-scale data sets

7.3.1 Dutch-to-English

To explore the question whether similar improvements to the ones reported in the previous section can be achieved with large-scale data sets, we carried out a similar series of experiments. Our first experimental data set is Dutch-to-English Europarl data (cf.

Table 17 Results on large-scale Dutch-to-English translation

Experiments	BLEU	NIST	METEOR	TER	WER	PER
Baseline	*27.29*	*6.686*	*56.81*	*58.65*	*63.97*	*45.18*
Words and part-of-speech tags						
Word±2	27.13	6.66	56.78	59.1	64.44	45.41
POS±2	26.93	6.67	56.51	59.06	64.19	45.51
POS±2†	26.9	6.67	56.61	58.94	64.10	45.52
Dependency relations						
PR	27.47 (66%)	6.726	57.02	58.5	63.59	45.10
OE	27.40	6.737	56.95	58.3	63.56	44.92
PR+OE	27.53 (63%)	6.721	57.15	58.64	63.93	45.08
PR+PW	27.17	6.690	56.86	58.94	64.09	45.34
PR+OE+PW	27.29	6.725	56.89	58.68	63.82	45.18
Combinations of words, part-of-speech tags and dependency relations						
PR+OE+Word±2	27.02	6.69	56.7	59.00	64.23	45.44
PR+OE+POS±2	27.14	6.64	56.68	59.39	64.56	45.76
PR+OE+POS±2†	27.16	6.66	56.58	59.00	64.17	45.53

Sect. 7.1). Analogous to the experiments on small-scale data sets, we experimented with adding contextual information features representing words, part-of-speech tags, dependency relations, and their combinations. We used the IGTree classifier to carry out these experiments, as TRIBL's memory needs become too demanding with data sets of this size.[15]

We used the same experimental settings as used with the small-scale Open Subtitles data set reported in Sect. 7.2.1. Experimental results are reported in Table 17, where we see that word and part-of-speech contexts are unable to yield any improvement over the Moses baseline. However, some of the dependency features produce small improvements over the baseline. Among the dependency features, PR+OE produces the largest improvement (0.24 BLEU points; 0.88% relative increase) over the baseline. However, none of the improvements are statistically significant with respect to the baseline score. Furthermore, we combine the best performing dependency feature combination (PR+OE) with Word±2, POS±2 and POS±2†, the results of which are shown in the last rows of the Table 17. Nevertheless, like lexical and POS features, none of the combinations are able to produce an improvement over the baseline. The other evaluation metrics tend to follow the trends of the BLEU evaluation metric.

In sum, word- and POS-based models do not show any improvements over a baseline PB-SMT model. In contrast, we achieve small but consistent improvements over the baseline for all evaluation metrics when dependency relations are employed as the source-language contextual feature.

[15] For example, the memory structure built by TRIBL takes about 90 GB when trained on a training set of 70 million instances generated on the 1.3 million training sentences dataset of English-to-Dutch, when only the Word±2 features are included.

Table 18 Results on English-to-Dutch translation employing homogeneous features

Experiments	BLEU	NIST	METEOR	TER	WER	PER
Baseline	*24.26*	*6.177*	*52.68*	*64.37*	*68.81*	*50.02*
Words and part-of-speech tags						
Word±2	24.50 (80%)	6.248	52.78	63.96	68.46	49.59
POS±2	24.04	6.150	52.17	64.44	68.69	50.1
Supertags						
CCG±1	24.58	6.229	52.46	63.79	68.2	49.85
LTAG±1	24.33	6.267	52.51	63.53	68.00	49.13
CCG±1 + LTAG±1†	24.45	6.250	52.54	63.87	68.30	49.74
Super-Pair±1†	24.35	6.184	52.31	64.45	68.71	50.32
Super-Pair±2†	24.14	6.160	52.34	64.56	68.92	50.31
CCG-LTAG±1	24.64 (96%)	6.235	52.79	63.90	68.27	49.58
Super-Pair±1	24.34	6.224	52.70	64.03	68.42	49.6
Dependency relations						
PR	24.72 (99.9%)	6.245	52.75	63.87	68.36	49.76
OE	24.32	6.219	52.62	64.00	68.52	49.87
PR+OE	24.62	6.235	52.82	63.95	68.26	49.81
PR+PW	24.58	6.260	52.80	63.62	68.33	49.49
PR+PW+OE	24.26	6.204	52.46	64.24	68.70	50.03
Semantic roles						
AL	24.56 (90.9%)	6.237	52.66	63.95	68.09	49.54
AL+PS	24.50 (91.6%)	6.221	52.50	64.15	68.33	49.79

7.3.2 English-to-Dutch

The first set of experiments to incorporate supertags on a large-scale data set were carried out on the same Dutch-to-English Europarl data set described in Sect. 7.3.1, but in the reverse direction. We also incorporated dependency features in this large-scale translation task as computed by the Malt dependency parser (cf. Sect. 7.2.2). Moreover, we introduce semantic roles as new contextual features in PB-SMT, and in addition we tried different combinations of lexical, syntactic and semantic features to test whether further improvements could be achieved.

Experimental results for individual contextual features are displayed in Table 18. As can be seen from the table, Word±2 yields a small (statistically insignificant) BLEU improvement (0.24 BLEU points, 0.98% relative increase) over the Moses baseline; POS±2 is unable to yield any improvement. Among the supertag-based experiments, CCG-LTAG±1 yields the highest improvement (0.38 BLEU points; 1.57% relative increase) over the baseline, which is statistically significant at the 96% level of confidence. Among the dependency features, PR produces the highest improvement (0.46 BLEU points; 1.90% relative increase) over the baseline, which is statistically significant at the 99.9% level of confidence. Among the semantic features, AL yields

Table 19 Results on English-to-Dutch translation combining best performing homogeneous features

Experiments	BLEU	NIST	METEOR	TER	WER	PER
Baseline	*24.26*	*6.177*	*52.68*	*64.37*	*68.81*	*50.02*
PR+Word±2	24.66 (92.9%)	6.302	52.89	63.36	67.95	49.09
PR+CCG-LTAG±1	24.51	6.301	52.55	63.14	67.32	49.04
PR+CCG-LTAG±1	24.55	6.232	52.58	63.72	68.01	49.48
AL+PR	24.70 (92.9%)	6.258	52.79	63.70	68.10	49.55
AL+CCG-LTAG±1	24.55	6.236	52.59	63.99	68.26	49.81
AL+PS+PR	24.72 (98.7%)	6.254	52.64	63.89	68.14	49.53
AL+PS+CCG-LTAG±1	24.50	6.218	52.77	64.23	68.35	49.73

the highest score (a 0.30 BLEU points improvement; 1.24% relative increase) over the baseline, but this is not statistically significant, although close to the significance level. Overall, PR remains the best performing feature among the individual context features.

Similar to the previous approaches, the best performing settings were combined to see whether further improvements could be achieved. Experimental results for the combined features are reported in Table 19. We see from Table 19 that a combined set-up (AL+PS+PR) equals the BLEU score obtained with the PR feature, and this improvement is statistically significant at the 98.2% level of confidence. Among the other combinations, Word±2 added to PR produces the second highest improvement (0.40 BLEU points, 1.64% relative increase) over the baseline, although this is not a statistically significant increase.

In sum, for English-to-Dutch translation (the reverse direction as compared to the previously reported experiment), improvements over the baseline for the dependency and supertag-based context-informed models are statistically significant in terms of BLEU at 99.9 and 96% levels of confidence respectively. In contrast, improvements for the word context are not statistically significant, and the POS-based model performs below the baseline PB-SMT model. Our novel semantic role contextual feature achieved modest gains over the baseline, both when used individually and in collaboration with other features. While this is encouraging, we note that semantic parsing is computationally expensive, so any gains in translation accuracy need to be offset against slower processing speed.

7.3.3 English-to-Japanese

Our next sets of experiments were carried out on a large-scale English-to-Japanese data set (cf. Sect. 7.1). As with the other large-scale experiments, the experiments were carried out using IGTree classifiers. Experimental results are shown in Table 20. None of the contextual features are able to improve on the Moses baseline with both test sets. Supertag-based features perform slightly better than POS and lexical features. As well as providing the largest amount of training data of all our experiments, the

Table 20 Experimental results for large-scale English-to-Japanese translation

Experiments	BLEU	NIST	TER	WER	PER
Baseline	*27.30*	*6.746*	*63.31*	*80.01*	*43.36*
Evaluation results on EJTestset1					
CCG±1	27.11	6.722	63.97	80.40	43.84
LTAG±1	27.18	6.736	63.53	80.06	43.51
CCG-LTAG±1	27.13	6.690	64.19	80.81	44.04
Super-Pair±1	27.10	6.727	63.84	80.44	43.59
POS±2	27.03	6.728	64.03	80.85	43.67
Word±2	26.65	6.656	64.18	80.20	43.83
Evaluation results on EJTestset2					
Baseline	*27.76*	*6.838*	*60.64*	*77.49*	*42.61*
CCG±1	27.41	6.768	61.49	78.38	43.23
LTAG±1	27.37	6.771	61.19	78.04	43.13
CCG-LTAG±1	27.31	6.734	61.68	78.51	43.27
Super-Pair±1	27.40	6.773	61.19	78.29	43.14
POS±2	27.39	6.744	61.65	78.79	43.18
Word±2	27.15	6.752	61.53	78.25	43.13

Table 21 Experimental results for large-scale English-to-Chinese translation

Experiments	BLEU	NIST	TER	WER	PER
Baseline	*9.85*	*4.87*	*77.13*	*82.03*	*60.29*
IGTree					
CCG±1	9.91	4.83	77.98	82.75	61.31
LTAG±1	9.80	4.82	77.71	82.58	61.12
TRIBL					
CCG±1	10.22 (99.3%)	4.96	76.68	81.65	59.76
LTAG±1	10.39 (99.9%)	4.89	76.92	81.82	60.20

NTCIR data has been reported to be very noisy (Okita et al. 2010), which may have affected the results.

7.3.4 English-to-Chinese

Our next set of experiments were carried out on an English-to-Chinese data set (cf. Sect. 7.1). Experiments were carried out with two types of supertags (CCG±1, LTAG±1), using both IGTree and TRIBL.

Experimental results are displayed in Table 21. When IGTree classifiers were used, LTAG±1 shows no improvement over the Moses baseline across all evaluation metrics, while CCG±1 shows a slight improvement on only the BLEU evaluation metric.

On the other hand, when we use TRIBL classifiers, CCG±1 yields a 0.37 BLEU points improvement (3.76% relative increase) over the baseline, a statistically

significant improvement at a 99.3% level of confidence. LTAG±1 produces the highest BLEU improvements (0.54 BLEU points; 5.48% relative increase) over the baseline, and the improvement is statistically significant at a 99.9% level of confidence.

In sum, while for large-scale English-to-Japanese translation none of the contextual features showed an improvement over the baseline, for large-scale English-to-Chinese translation a slight improvement over the baseline in terms of BLEU was observed for the CCG supertag context when IGTree is used for the classification task. We also carried out experiments using TRIBL as the classifier, and achieved significant improvements over the baseline for both the CCG and LTAG supertag features.

Comparing the effectiveness of the classifiers on large-scale translation tasks, IGTree proved useful for Dutch-to-English and English-to-Dutch, but not for English-to-Japanese or English-to-Chinese. In contrast, TRIBL was effective for the latter language direction. In terms of contextual features, overall supertags, dependency relations and semantic roles seemed to be more effective than word- and POS-based models.

7.4 Learning curve experiments

7.4.1 English-to-Spanish

Thus far, combining two different approximate memory-based classifiers (of which TRIBL has empirically tuned hyperparameters), different data set sizes, different language pairs, text genres and domains, various source-side context features and context widths, we have obtained a mixed bag of results. We observe that the context-informed models tend to perform better than the baseline PB-SMT model on small-scale training data, but the relative gain tends to diminish when we use larger training sets, which may be partly due to the approximate behaviour of the IGTree classifier compared to unrestriced k-nearest neighbour classification. With the English-to-Chinese experiments we observed that the improvements may be somewhat larger with the TRIBL classifier, a closer approximation of k-nearest neighbour classification, than with the faster IGTree algorithm. So far, however, we have not systematically varied the amount of training data sizes given a particular data set, to see whether the relative advantage of TRIBL over IGTree changes with the amount of training data available, and how their performance relates to the baseline with varying amounts of training data available.

In this section we explore a new language pair, English-to-Spanish. We conduct *learning curve* experiments on increasing training data sets while adding an optimized set of contextual features. We segment the English-to-Spanish data set into several incremental slices of increasing size, and perform a series of experiments on each of these data sets.

To conduct learning curve experiments, we employ the Spanish-to-English data set (cf. Sect. 7.1). We segmented the English-to-Spanish training set into eight pseudo-exponentially increasing training sets: 10 K, 20 K, 50 K, 100 K, 200 K, 500 K, 1 M, and 1,639,764 training sentences. To perform experiments on this sequence of training sets, we used both IGTree and TRIBL. We were only able to use the TRIBL classifier

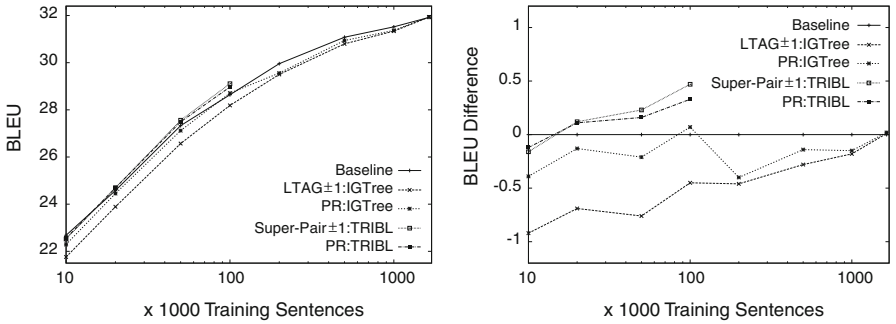

Fig. 7 BLEU learning curves (*left*) and difference curves (*right*) comparing the Moses baseline against two IGTree (LTAG±1 and PR) and TRIBL (Super-Pair±1 and PR) classifiers

with training sets containing up to 100 K sentences due to TRIBL's relatively high memory requirements.

We plot the BLEU score learning curves of the two best-performing context-informed models for TRIBL (Super-Pair±1 and PR) and for IGTree (LTAG±1 and PR) as well as the Moses baseline in the left part of Fig. 7. The figure adopts a logarithmic horizontal axis, representing the number of training sentences. In addition, the right-hand side graph of Fig. 7 shows the BLEU difference curves of the four classifier experiments against the baseline, highlighting the gains and losses against the baseline. The curves of IGTree extend up to the maximum of 1.64 M training sentences; as noted, due to limitations in memory, the TRIBL experiment extends to up to 100 K training sentences.

Figure 7 shows that the Super-Pair±1 and PR curves of TRIBL start just under the baseline curve, then increasingly improve over themselves and the baseline curve. The figure also illustrates that the LTAG±1 and PR curves of IGTree start at a lower level than the baseline curve, and end at the same level as the baseline curve at the largest training set size.

To summarize, (a) TRIBL appears to be effective on both small and large-scale data sets, though its memory needs prohibit it from being used with the largest-sized training sets; on the other hand, it does not improve the context-informed models on the smallest amounts of training data tested (e.g. 10 K sentences); (b) IGTree does not offer improvements over the baseline either with the small or the large-scale context-informed models; the performance of the large-scale context-informed models with the IGTree classifier are merely close or equal to the performance of the Moses baseline.

As an additional point of analysis, Fig. 8 compares the Moses baseline with both TRIBL and IGTree using the CCG±1 feature, in terms of the average number of target phrases considered for a source phrase for varying training data sizes. The CCG±1 feature produces a similar performance to LTAG±1 and Super-Pair±1 when IGtree and TRIBL classifiers are used, respectively.

The graph in Fig. 8 shows that the TRIBL curve lies between the IGTree curve and the Moses baseline curve; the Moses baseline uses an increasing number of target phrases with more training data, reaching an average of several hundreds of phrases at

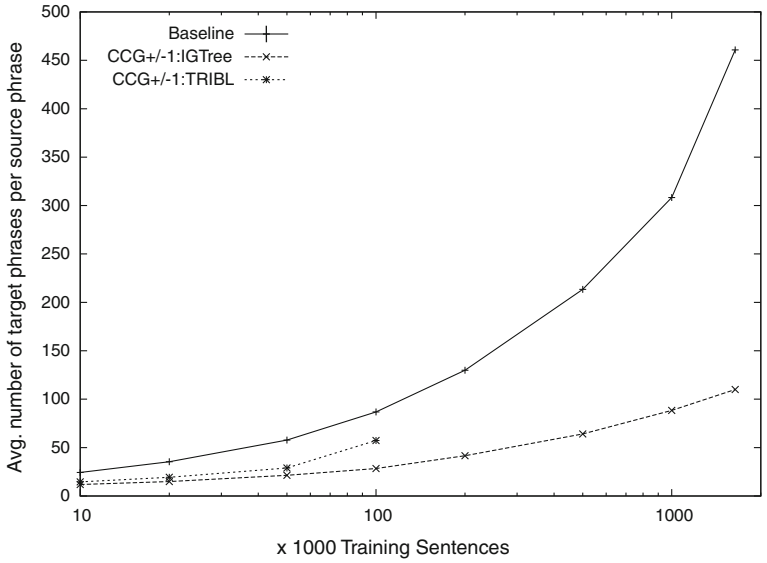

Fig. 8 Average number of target phrase distribution sizes for source phrases for TRIBL and IGTree compared to the Moses baseline

the maximal training set sizes. The TRIBL curve starts close to the IGTree curve, but rises at 100 K training sentences; nevertheless, the TRIBL curve remains under the baseline curve. Thus, both the TRIBL and IGTree classifiers produce smaller, more constrained distributions of the target phrases given a source phrase and its context information.

Expanding on the results displayed in Fig. 7, we analyse four specific learning curves for different context-informed models using the TRIBL classifier. Figure 9 visualizes the losses and gains in terms of BLEU score of experiments with four individual types of contextual features compared against the baseline: using supertags (Super-Pair±1), dependency relations (PR), POS tags, and words. The figure shows that the Super-Pair±1, POS, PR and word curves start below the baseline curve at the smallest training set size (10,000 sentences), but then start to deviate positively and increasingly from the baseline curve when more training data is added.

7.4.2 Dutch-to-English

Originally we ran our large-scale experiments on the Dutch-to-English and English-to-Dutch language pairs with the IGTree classifiers (cf. Sect. 7.3). As mentioned earlier, we observe in the English-to-Chinese translation task (cf. Sect. 7.3.4) and the English-to-Spanish learning curve experiments (cf. Sect. 7.4.1) that TRIBL seems to be a more effective classifier than IGTree in improving the performances of the context-informed SMT systems, but we were able to use TRIBL classifiers only for the small-scale translations due to its relatively high memory requirements. However,

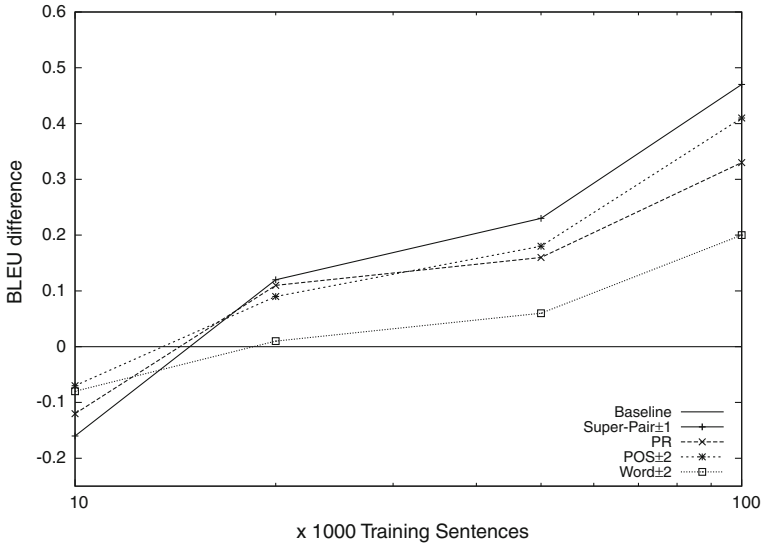

Fig. 9 BLEU difference curves of four context-informed models using TRIBL

TRIBL was reprogrammed recently, and can now efficiently handle large number of examples. This inspires us to deploy TRIBL classifiers for the large-scale translation.

To investigate the consequences of the different context-informed SMT systems on the increasing sizes of training sets while employing TRIBL as the classifier, we carried out experiments on the Dutch-to-English and English-to-Dutch language pairs. First, like the division of English-to-Spanish data set (cf. Sect. 7.4.1), we segmented the Dutch-to-English training set (cf. Sect. 7.1) into eight pseudo-exponentially increasing training sets: 10 K, 20 K, 50 K, 100 K, 200 K, 500 K, 1 M, and 1,311,111 training sentences. In this section, we report the outcomes of the Dutch-to-English learning curve experiments.

In order to perform the learning curve experiments, we chose the previously best-performing experimental set-ups for each feature type (dependency relations (PR, OE), part-of-speech tags (POS±2), and neighbouring words (Word±2)). Figure 10 shows the learning curves and score-difference curves comparing the Dutch-to-English Moses baseline against the four context-based models (PR, OE, POS±2, Word±2). The three left-hand side graphs in Fig. 10 show respectively BLEU (top), METEOR (centre) and TER (bottom) learning curves representing the performance of four context-informed models compared against the baseline. The three right-hand side graphs in Fig. 10 show respectively BLEU (top), METEOR (centre) and TER (bottom) score-difference curves, highlighting the gains and losses against the baseline.

We observe that the BLEU and METEOR curves (learning and score-difference) of all the context-informed models always remain above the baseline curve from the starting point (10 K training data) to the end point (1.31 M training data). Figure 10 also shows that the PR and OE learning curves (BLEU and METEOR) start at a lower level than the POS±2 and Word±2 curves at smaller training set sizes (10–500 K training set), but end at the same level as the Word±2 curve, and at a higher level

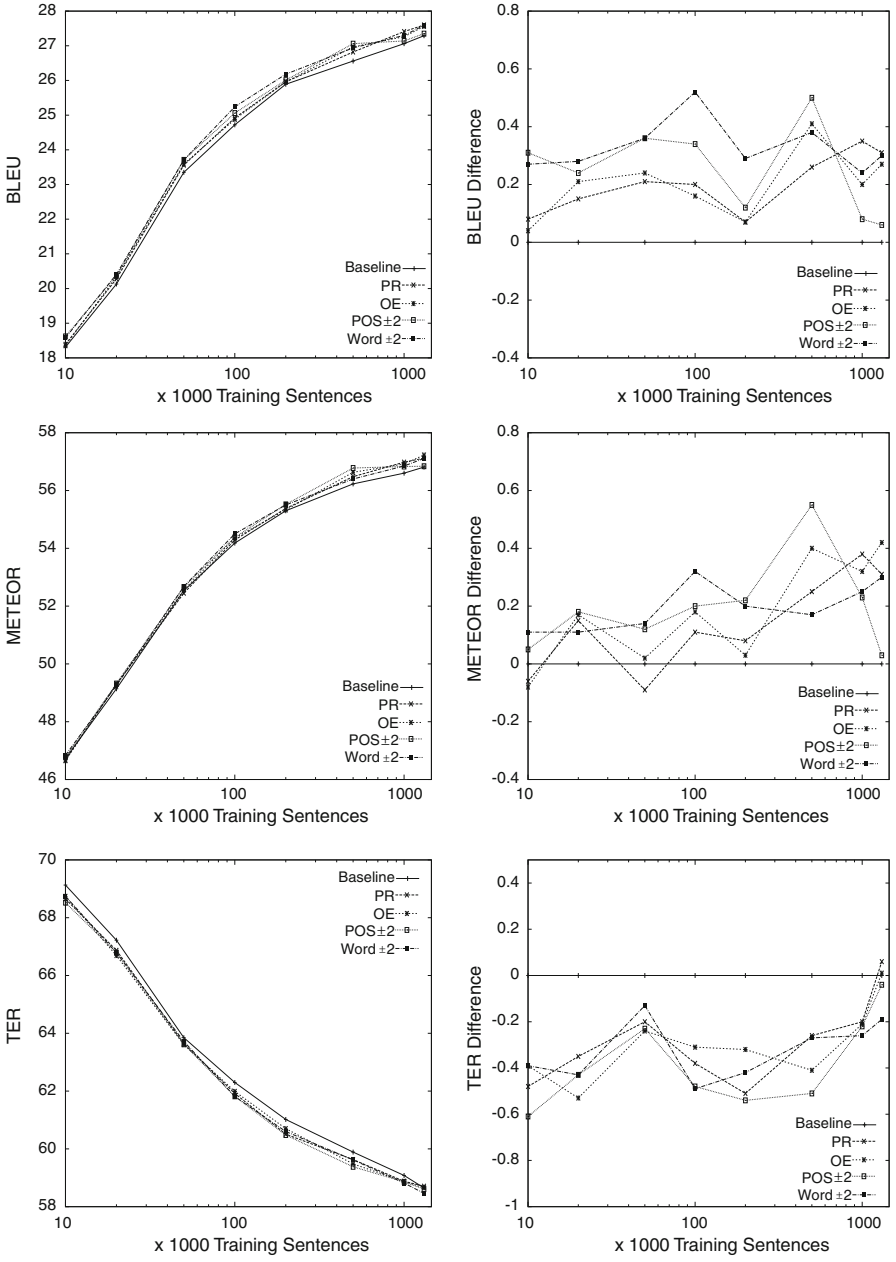

Fig. 10 Dutch-to-English Learning curves (*left-hand side graphs*) and difference curves (*right-hand side graphs*) comparing the Moses baseline against four context-informed models (PR, OE, POS±2 and Word±2). These curves are plotted with scores obtained using three evaluation metrics: BLEU (*top*), METEOR (*centre*) and TER (*bottom*)

than the POS±2 curve for the larger training set sizes (1–1.31 M training set). Note that TER is an error metric, so lower scores indicate better performance. The bottom two graphs of Fig. 10 show that the TER curves (learning and score-difference) of all context-informed models (PR, OE, POS±2, Word±2) mostly remain below the baseline curve, which indicates the effectiveness of source-language context in this translation task also in terms of TER.

In summary, in the Dutch-to-English translation task the dependency relations and neighbouring words appear to be more effective source-language context features than POS tags according to the performance measured by all evaluation metrics.

7.4.3 English-to-Dutch

In this section, we report the outcomes of the English-to-Dutch learning curve experiments. In order to conduct English-to-Dutch learning curve experiments we consider the previously best-performing experimental set-ups comprising each feature type: supertags (CCG±1, LTAG±1), dependency relations (PR), semantic roles (PS-AL), and basic context features (POS±2, Word±2). Figure 11 illustrates BLEU (top), METEOR (centre) and TER (bottom) learning curves (left-hand side graphs) and score-difference curves (right-hand side graphs) comparing the Moses baseline against the six context-informed models (CCG±1, LTAG±1, PR, PS-AL, POS±2 and Word±2).

We see from the top-left and -right graphs in Fig. 11 that the semantic and dependency feature-based BLEU curves (PS-AL, PR) show consistency in residing mostly above the baseline BLEU curve from the starting point (10 K training data) to the end point (1.31 M training data). Supertag-based BLEU learning curves (CCG±1, LTAG±1) start close to the baseline curve, go upwards and cross the baseline curve while adding more training data. Interestingly, LTAG±1 and CCG±1 produced respectively the highest and second highest BLEU improvements over the Moses baseline for the larger amounts of training data. We observed that most of the improvements over the baseline with supertag context features are statistically significant. In contrast, we found that the most of the improvements for word and POS-based models over the baseline are not statistically significant. In short, supertags appear to be the most effective context features in PB-SMT according to the performance measured by the BLEU evaluation metric.

Centre-left and -right graphs of Fig. 11 show METEOR learning and score-difference curves, respectively. We see that the most of the METEOR curves (learning and score-difference) do not resemble the BLEU curves (top-left and -right graphs in Fig. 11). The PS-AL and PR-based METEOR curves show consistency in residing mostly above the baseline curve for all amounts of training data, while the Word ± 2-based METEOR curve shows consistency in residing above the baseline curve at larger amounts (500 K, 1 M and 1.31 M) of training data. Interestingly, the supertag- and POS-based METEOR curves reside mostly beneath the baseline curve.

The bottom-left and -right graphs in Fig. 11 show respectively TER learning and score-difference curves. We see from the bottom-part of the Fig. 11 that most of the TER learning and score-difference curves show consistency in residing below the baseline. Interestingly, both LTAG±1 and CCG±1 produce higher TER scores than

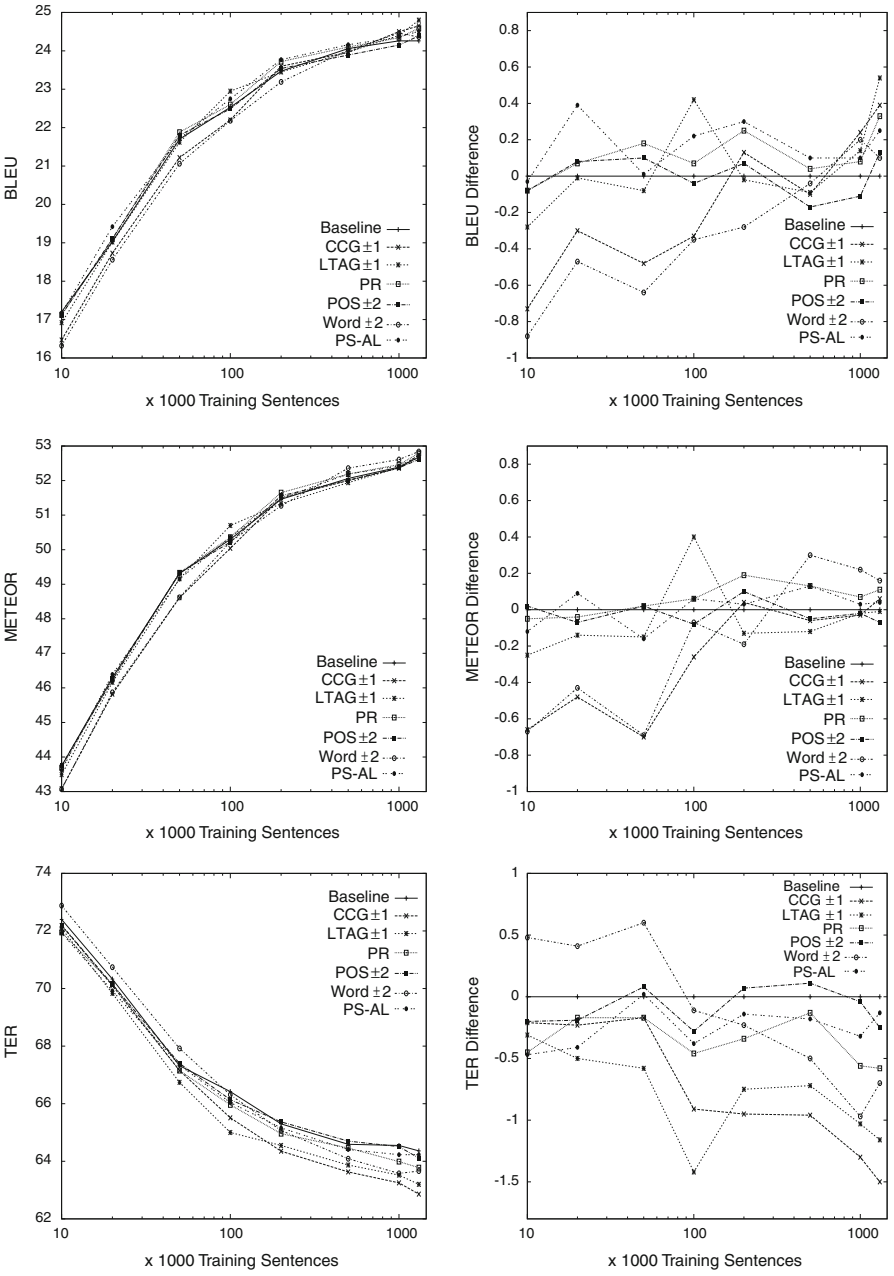

Fig. 11 English-to-Dutch Learning curves (*left-hand side graphs*) and difference curves (*right-hand side graphs*) comparing the Moses baseline against four context-informed models (CCG±1, LTAG±1, PR, PS-AL, POS±2 and Word±2). These curves are plotted with scores obtained using three evaluation metrics: BLEU (*top*), METEOR (*centre*) and TER (*bottom*)

the baseline and other context-informed models (PR, PS-AL, POS±2, Word±2) at all amounts of training data.

In summary, in this translation task, dependency (PR) and semantic (PS-AL) features appear to be the most effective source-language context features in PB-SMT according to the performance measured by all evaluation metrics. Nevertheless, the BLEU and TER metrics point at supertags (CCG and LTAG) as the most effective context features, while the METEOR evaluation metric suggests otherwise. As the METEOR metric does not support the Dutch language (Lavie and Agarwal 2007), we operated it with its default English settings. We suspect this might be the reason why the METEOR shows inconsistency while evaluating the Dutch sentences.

7.5 Translation examples

We performed manual qualitative analysis comparing the translated outputs of the best-performing systems with those of the Moses baseline systems. In order to carry out the manual evaluation, we randomly sampled fifty (50) test set sentences from the translated output of each system.

7.5.1 Dutch-to-English translation examples

First, we looked at the translated output of our best-performing system (PR+OE+POS ±2†) against that of the Moses baseline in the small-scale Dutch-to-English translation task (cf. Sect. 7.2.1). We observed that the (PR+OE+POS±2†) system generates a more fluent as well as more adequate output than the baseline for most sentences. The following are two translation examples:

(5)	Dutch:	heeft mijn vader je gestuurd ?
	Reference:	did my father send you ?
	PR+OE+POS±2†:	*did* my father *send* you ?
	Baseline:	my father *has sent* you ?
(6)	Dutch:	daarna vraag je om informatie.
	Reference:	then you call for information.
	PR+OE+POS±2†:	then *you ask* for information.
	Baseline:	then *ask you* to get information.

In Example (5) we observe that the baseline does not select the word order of a question, and selects a less optimal (more literal) translation of the Dutch auxiliary verb *heeft, has*, while the PR+OE+POS system generates a translation identical to the reference translation. In example (6), the baseline system again selects a less appropriate (Dutch) word order.

7.5.2 English-to-Dutch translation examples

We also analysed the translated output of our best-performing system (LTAG±1) against that of the Moses baseline in the English-to-Dutch translation task (cf. Sect. 7.4.3). We observed that the baseline Moses system frequently mistranslates

English function words. The following are the two translation examples which illustrate how our best-performing system (LTAG±1) surpasses the Moses baseline in this translation task:

(7)	English:	European agriculture is not uniform.
	Reference:	De Europese landbouw is verre van eenvormig.
	LTAG±1:	De Europese landbouw is niet uniform.
	Baseline:	Europese landbouw niet uniform is.
(8)	Baseline:	Apart from a limited budget, the European Union has little political interest in Tajikistan.
	Reference:	Naast een beperkte begroting heeft de Europese Unie politiek gezien weinig in Tadzjikistan te zoeken.
	LTAG±1:	Afgezien van een beperkte begroting heeft de Europese Unie weinig politieke belang in Tadzjikistan.
	Baseline:	Afgezien van een beperkte begroting, de Europese Unie heeft weinig politieke belang in Tadzjikistan.

In the translation example (7), the translation produced by LTAG±1 is fluent and roughly synonymous to the reference translation, while the baseline generates a translation with a wrong word order, and also misses the initial article 'De'. Translation example (8) resembles (7) in that the LTAG±1 system generates a fluent and grammatical translation save for one agreement issue (the adjective *politieke* should be *politiek* as the noun *belang* has neuter gender), while the baseline system also generates a faulty word order.

8 Conclusions and future work

In this paper, we presented a revised, extended account of our previous work on using a range of features as source-language context to better enable a state-of-the-art PB-SMT system to select appropriate target language phrases for consideration in the generation of the most probable translation given the input. Such features include neighbouring position-specific lexical and part-of-speech features of words surrounding the phrase to be translated, as well as information linking the head word of the phrase to its syntactic context in terms of supertags or dependency relations.

While parts of this appeared in previously published research (Haque et al. 2009a,b), in this paper we added a number of novel aspects, including using semantic roles as new contextual features in PB-SMT, adding new language pairs, and examining the scalability of our research to larger amounts of training data.

The most significant improvements observed in our experiments involve the integration of long-distance contextual features, such as dependency relations in combination with part-of-speech tags in Dutch-to-English subtitle translation, the combination of dependency parse and semantic role information in English-to-Dutch parliamentary debate translation, or CCG and LTAG supertag features in English-to-Chinese translation.

As far as scalability is concerned, when our PB-SMT systems were trained with larger amounts of parallel data, the effects of the source-language context are lessened somewhat, but in some cases remain statistically significant. For English-to-Dutch, for example, while the POS-based model failed to contribute positively, our dependency- and supertag-based improvements continued to be effective. Furthermore, our novel use of semantic roles as a source-language discriminative feature showed encouraging improvements over the PB-SMT baseline.

When varying the amounts of English-to-Spanish Europarl training data used from 10,000 to 1.64 million sentences in a learning curve experiment, the resulting curves demonstrate that gains attained by our source-language contextual models cannot be expected to occur given any amount of training data. We observe that the TRIBL classifier attains gains at small training set sizes, though not at the smallest sizes (10,000 training sentences). IGTree, on the other hand, disappoints by requiring the maximal amount of training data (1.64 million sentences) to equal the baseline. Furthermore, learning curve experiments on the Dutch-to-English and English-to-Dutch language pairs show that rich and complex syntactic features surpass basic features (words and POS tags) as source-language context features in the small-scale as well as the large-scale translations. Moreover, outcomes of the manual analysis conducted on the MT outputs of the several context-informed models against the respective Moses baselines justify our claims established on the basis of the gains obtained with several automatic evaluation measures. We argue that, in general, learning curve experiments give a more complete overview of relative gains when more data is available. For attaining higher performance, using more training data remains the best advice.

To summarize our findings, we have shown that whatever language pair might need to be deployed, using source-language context is guaranteed to produce better translations compared to a baseline PB-SMT system. To be more precise, if one has a parser available for the source language at hand, integrating syntactic dependency information pertaining to the current input string can generate improved translation quality. Alternatively, if no such parser is available, then POS or supertag information can be useful, but if even this is absent, then taking the neighbouring words into account is also likely to be effective. Such source-language contextual models become less effective when scaling to large amounts of parallel data, yet even here, statistically significant scores are still to be seen. Furthermore, our experiments have been carried out on a wide range of language pairs, and on a variety of domains of training material.

As for future work, we aim to conduct a suite of experiments to investigate the effectiveness of our models on state-of-the-art hierarchical phrase-based SMT systems (Chiang 2007); our initial effort in this direction (Haque et al. 2010) is encouraging. In addition, apart from experimenting with still more language pairs and different types of training data, we would like to provide a comprehensive guide on how best to combine different source-language contextual features where more than one type is available, and if possible, to predict a priori—perhaps based on the combination of language pair and training data type—the optimal features to use in such circumstances. As a first step we would investigate the influence of the degree to which a domain triggers formulaic language. If a domain contains largely formulaic language, selecting only simple lexical features such as neighbouring words could already be effective, as the generalizing power of more abstract linguistic features is not needed. The reverse may

be the case in more open, less formulaic domains. Our memory-based classifiers could be used to provide a quantitative estimate of how similar unseen sequences are to training sentences, much like a fuzziness score in translation memories or example-based machine translation.

Acknowledgements This work is supported by Science Foundation Ireland (grant no. 07/CE/ I1142) and the Irish Centre for High-End Computing.[16] The work of Van den Bosch is supported by the Netherlands Organisation for Scientific Research (NWO) as part of the "Implicit Linguistics" Vici project, which also provided computing infrastructure. We would like to thank Yifan He and Sergio Penkale for their input on the presentation of translation examples in Chinese and Spanish languages, respectively.

References

Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Mach Learn 6(1):37–66

Bangalore S, Joshi AK (1999) Supertagging: an approach to almost parsing. Comput Linguist 25(2):237–265

Bangalore S, Haffner P, Kanthak S (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 2007), Prague, Czech Republic, pp 152–159

Berger AL, Della Pietra VJ, Della Pietra SA (1996) A maximum entropy approach to natural language processing. Comput Linguist 22(1):39–71

Brown PF, Cocke J, Della Pietra SA, Della Pietra VJ, Jelinek F, Lafferty JDD, Mercer RL, Roossin PS (1990) A statistical approach to machine translation. Comput Linguist 16(2):79–85

Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL (1991) A statistical approach to sense disambiguation in machine translation. In: Proceedings of the workshop on speech and natural language, HLT 1991, Pacific Grove, CA, pp 146–151

Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL (1993) The mathematics of statistical machine translation: parameter estimation. Comput Linguist 19(2):263–311

Brunning J, Gispert A, Byrne W (2009) Context-dependent alignment models for statistical machine translation. In: NAACL HLT 2009: proceedings of human language technologies: the 2009 annual conference of the North American chapter of the ACL, Boulder, CO, pp 110–118

Carpuat M, Wu D (2005) Word sense disambiguation vs. statistical machine translation. In: 43rd Annual meeting of the association for computational linguistics (ACL 2005), University of Michigan, Ann Arbor, MI, pp 387–394

Carpuat M, Wu D (2007) Improving statistical machine translation using word sense disambiguation. In: EMNLP-CoNLL-2007: proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning, Prague, Czech Republic, pp 61–72

Carreras X, Márquez L (2004) Introduction to the CoNLL-2004 shared task: semantic role labeling. In: Proceedings of the CoNLL 2004 shared task, Boston, MA, pp 89–97

Chen J, Bangalore S, Vijay-Shanker K (2006) Automated extraction of tree-adjoining grammars from treebanks. Nat Lang Eng 12(3):251–299

Chan YS, Ng HT, Chiang D (2007) Word sense disambiguation improves statistical machine translation. In: Proceedings of the 45th annual meeting of the association for computational linguistics (ACL 2007), Prague, Czech Republic, pp 33–40

Chiang D (2007) Hierarchical phrase-based translation. Comput Linguist 33(2):202–228

Chiang D, Knight K, Wang W (2009) 11,001 new features for statistical machine translation. In: Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the association for computational linguistics (HLT NAACL 2009), Boulder, CO, pp 218–226

Clark S, Curran JR (2004) The importance of supertagging for wide-coverage CCG parsing. In: Proceedings of the 20th international conference on computational linguistics (COLING 2004), Geneva, Switzerland, pp 282–288

[16] http://www.ichec.ie.

Daelemans W, van den Bosch A (2005) Memory-based language processing. Cambridge University Press, Cambridge

Daelemans W, van den Bosch A, Weijters A (1997) IGTree: using trees for compression and classification in lazy learning algorithms. Artif Intell Rev 11:407–423

Daelemans W, van den Bosch A, Zavrel J (1997b) A feature-relevance heuristic for indexing and compressing large case bases. In: Van Someren M, Widmer G (eds) Poster papers of the ninth European conference on machine learning, Prague, Czech Republic, pp 29–39

Doddington G (2002) Automatic evaluation of language translation using n-gram cooccurrence statistics. In: HLT 2002: human language technology conference: proceedings of the second international conference on human language technology research, San Diego, CA, pp 138–145

Foster G, Kuhn R, Johnson H (2006) Phrasetable smoothing for statistical machine translation. In: EMNLP-2006: proceedings of the 2006 conference on empirical methods in natural language processing, Sydney, Australiapages, pp 53–61

Galley M, Graehl J, Knight K, Marcu D, DeNeefe S, Wang W, Thayer I (2006) Scalable inference and training of context-rich syntatic translation models. In: Coling-ACL 2006: proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics, Sydney, Australia, pp 961–968

García-Varea I, Och FJ, Ney H, Casacuberta F (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 the association for computational linguistics (ACL/EACL 2001), Toulouse, France, pp 204–211

García-Varea I, Och FJ, Ney H, Casacuberta F (2002) Improving alignment quality in statistical machine translation using context-dependent maximum entropy models. In: Proceedings of the 19th international conference on computational linguistics (Coling 2002), Taipei, Taiwan, pp 1051–1054

Giménez J, Màrquez L (2007) Context-aware discriminative phrase selection for statistical machine translation. In: Proceedings of the second workshop on statistical machine translation, ACL 2007, Prague, Czech Republic, pp 159–166

Giménez J, Màrquez L (2009) Discriminative phrase selection for statistical machine translation. In: Goutte C, Cancedda N, Dymetman M, Foster G (eds) Learning machine translation. NIPS Workshop Series. MIT Press, Cambridge

Gimpel K, Smith NA (2008) Rich source-side context for statistical machine translation. In: Proceedings of the third workshop on statistical machine translation, ACL-08:HLT, Columbus, OH, pp 9–17

Gimpel K, Smith NA (2009) Feature-rich translation by quasi-syncronous lattice parsing. In: EMNLP-2009: proceedings of the 2009 conference on empirical methods in natural language processing, Singapore, pp 219–228

Haque R, Naskar SK, Ma Y, Way A (2009a) Using supertags as source language context in SMT. In: EAMT-2009: proceedings of the 13th annual conference of the European association for machine translation, Barcelona, Spain, pp 234–241

Haque R, Naskar SK, van den Bosch A, Way A (2009b) Dependency relations as source context in phrase-based SMT. In: Proceedings of PACLIC 23: the 23rd pacific asia conference on language, information and computation, Hong Kong, China, pp 170–179

Haque R, Naskar SK, van den Bosch A, Way A (2010) Supertags as source language context in hierarchical phrase-based SMT. In: Proceedings of AMTA 2010: the ninth conference of the association for machine translation in the Americas, Denver, CO, pp 210–219

Hasan S, Ganitkevitch J, Ney H, Andrés-Ferrer J (2008) Triplet lexicon models for statistical machine translation. In: EMNLP 2008: Proceedings of the 2008 conference on empirical methods in natural language processing, Honolulu, HI, pp 372–381

Hockenmaier J (2003) Data and models for statistical parsing with combinatory categorial grammar. PhD thesis, University of Edinburgh, UK

Ittycheriah A, Roukos S (2007) Direct translation model 2. In: NAACL-HLT-2007 human language technology: the conference of the North American chapter of the association for computational linguistics, Rochester, NY, pp 57–64

Johansson R, Nugues P (2008) Dependency-based syntactic-semantic analysis with PropBank and NomBank. In: Proceedings of the CoNLL-2008 shared task, Manchester, UK, pp 183–187

Koehn P (2004a) Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In: Frederking Robert E, Taylor Kathryn B (eds) Machine translation: from real users to research: 6th

conference of the association for machine translation in the Americas, AMTA 2004, Washington, DC, pp 115–124

Koehn P (2004b) Statistical significance tests for machine translation evaluation. In: EMNLP-2004: Proceedings of the 2004 conference on empirical methods in natural language processing, Barcelona, Spain, pp 388–395

Koehn P (2005) Europarl: a parallel corpus for statistical machine translation. In: MT summit X, the tenth machine translation summit, Phuket, Thailand, pp 79–86

Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. In: HLT-NAACL 2003: conference combining human language technology conference series and the North American chapter of the association for computational linguistics conference series, Edmonton, AB, pp 48–54

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Proceedings of the demo and poster sessions, ACL 2007, Prague, Czech Republic, pp 177–180

Lavie A, Agarwal A (2007) METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the second workshop on statistical machine translation, ACL 2007, Prague, Czech Republic, pp 228–231

Liang P, Bouchard-Côté A, Klein D, Taskar B (2006) An end-to-end discriminative approach to machine translation. In: Coling-ACL 2006: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics, Sydney, Australia, pp 761–768

Marton Y, Resnik P (2008) Soft syntactic constraints for hierarchical phrased-based translation. In: Proceedings of the 46th annual meeting of the association for computational linguistics: human language technologies (ACL-08: HLT), The Ohio State University, Columbus, OH, pp 1003–1011

Mauser A, Hasan S, Ney H (2009) Extending statistical machine translation with discriminative and trigger-based Lexicon models. In: EMNLP-2009: proceedings of the 2009 conference on empirical methods in natural language processing, Singapore, pp 210–218

Max A, Makhloufi R, Langlais P (2008) Explorations in using grammatical dependencies for contextual phrase translation disambiguation. In: EAMT 2008: 12th annual conference of the European association for machine translation, Hamburg, Germany, pp 114–119

Nivre J, Hall J, Nilsson J (2006) MaltParser: a data-driven parser generator for dependency parsing. In: LREC 2006: Proceedings of the fifth international conference on language resources and evaluation, Genoa, Italy, pp 2216–2219

Och FJ (2003) Minimum error rate training in statistical machine translation. In: 41st Annual meeting of the association for computational linguistics (ACL 2003), Sapporo, Japan, pp 160–167

Och FJ, Ney H (2000) A comparison of alignment models for statistical machine translation. In: Coling 2000: the 18th international conference on computational linguistics, Saarbrücken, Germany, pp 1086–1090

Och FJ, Ney H (2002) Discriminative training and maximum entropy models for statistical machine translation. In: 40th Annual meeting of the association for computational linguistics (ACL 2002), Philadelphia, PA, pp 295–302

Okita S, Jiang J, Haque R, Al-Maghout H, Du J, Naskar SK, Way A (2010) MaTrEx: the DCU MT system for NTCIR-8. In: Proceedings of NTCIR-8, Tokyo, Japan, pp 377–383

Papineni K, Roukos S, Zhu W (2002) BLEU: a method for automatic evaluation of machine translation. In: 40th Annual meeting of the association for computational linguistics (ACL 2002), Philadelphia, PA, pp 311–318

Patry A, Langlais P (2009) Prediction of words in statistical machine translation using a multilayer perceptron. In: MT Summit XII: proceedings of the twelfth machine translation Summit, Ottawa, ON, Canada, pp 101–111

Penkale S, Haque R, Dandapat S, Banerjee P, Srivastava AK, Du J, Pecina P, Naskar SK, Forcada ML, Way A (2010) MATREX: the DCU MT system for WMT 2010. In: Proceedings of the joint fifth workshop on statistical machine translation and metrics MATR (WMT-MetricsMATR 2010), ACL 2010, Uppsala, Sweden, pp 143–148

Quirk C, Menezes A, Cherry C (2005) Dependency treelet translation: syntactically informed phrasal SMT. In: ACL-2005: 43rd annual meeting of the association for computational linguistics, Ann Arbor, MI, pp 271–279

Shen L, Zhang B, Matsoukas S, Weischedel R (2009) Effective use of linguistic and contextual information for statistical machine translation. In: EMNLP-2009: proceedings of the 2009 conference on empirical methods in natural language processing, Singapore, pp 72–80

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: AMTA 2006: Proceedings of the 7th Conference of the association for machine translation in the Americas, Cambridge, MA, pp 223–231

Specia L, Sankaran B, Nunes MGV (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 2008), Haifa, Israel, pp 399–410

Steedman M (2000) The syntactic process. MIT Press, Cambridge, MA

Stroppa N, van den Bosch A, Way A (2007) Exploiting source similarity for SMT using context-informed features. In: Proceedings of the 11th international conference on theoretical and methodological issues in machine translation (TMI 2007), Skövde, Sweden, pp 231–240

Surdeanu M, Johansson R, Meyers A, Màrquez L, Nivre J (2008) The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In: Proceedings of the 12th conference on computational natural language learning (CoNLL-2008), Manchester, UK, pp 159–177

Tiedemann J, Nygaard L (2004) The OPUS corpus—parallel & free. In: Proceedings of the 4th international conference on language resources and evaluation (LREC 2004), Lisbon, Portugal, pp 1183–1186

Tillmann C, Zhang T (2006) A discriminative global training algorithm for statistical mt. In: Coling-ACL 2006: proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics, Sydney, Australia, pp 721–728

van den Bosch A (2004) Wrapped progressive sampling search for optimizing learning algorithm parameters. In: Verbrugge R, Taatgen N, Schomaker L (eds) Proceedings of the 16th Belgian-Dutch conference on artificial intelligence, Groningen, The Netherlands

van den Bosch A, Busser B, Canisius S, Daelemans W (2007) An efficient memorybased morpho-syntactic tagger and parser for Dutch. In: Proceedings of computational linguistics in the Netherlands: selected papers from the seventeenth CLIN meeting, Leuven, Belgium, pp 99–114

Venkatapathy S (2008) NLP tools contest—2008: summary. In: Proceedings of the NLP tools contest, ICON 2008, Pune, India

Venkatapathy S, Bangalore S (2007) Three models for discriminative machine translation using global lexical selection and sentence reconstruction. In: SSST, NAACL-HLT-2007 AMTA workshop on syntax and structure in statistical translation, Rochester, NY, pp 96–102

Vickrey D, Biewald L, Teyssier M, Koller D (2005) Word-sense disambiguation for machine translation. In: HLT-EMNLP-2005: proceedings of human language technology conference and conference on empirical methods in natural language processing, Vancouver, BC, Canada, pp 771–778

Wu D, Fung P (2009) Can semantic role labeling improve SMT?. In: EAMT-2009: proceedings of the 13th annual conference of the European association for machine translation, Barcelona, Spain, pp 218–225

Xiong D, Zhang M, Li H (2010) Learning translation boundaries for phrase-based decoding. In: NAACL-HLT-2010: human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics, Los Angeles, CA, pp 136–144

Zens R, Ney H (2004) Improvements in phrase-based statistical machine translation. In: HLT-NAACL 2004: human language technology conference and North American chapter of the association for computational linguistics annual meeting, Boston, MA, pp 257–264