Statistical Machine Translation Enhancements through Linguistic Levels: A Survey

MARTA R. COSTA-JUSSÀ, Institute for Infocomm Research, Singapore
MIREIA FARRÚS, Universitat Pompeu Fabra, Barcelona

Machine translation can be considered a highly interdisciplinary and multidisciplinary field because it is approached from the point of view of human translators, engineers, computer scientists, mathematicians, and linguists. One of the most popular approaches is the Statistical Machine Translation (SMT) approach, which tries to cover translation in a holistic manner by learning from parallel corpus aligned at the sentence level. However, with this basic approach, there are some issues at each written linguistic level (i.e., orthographic, morphological, lexical, syntactic and semantic) that remain unsolved. Research in SMT has continuously been focused on solving the different linguistic levels challenges. This article represents a survey of how the SMT has been enhanced to perform translation correctly at all linguistic levels.

Categories and Subject Descriptors: I.2.7

General Terms: Survey

Additional Key Words and Phrases: Linguistics, statistical machine translation, orthography, morphology, Lexis, syntax, semantics

ACM Reference Format:
Marta R. Costa-jussà and Mireia Farrús. 2014. Statistical machine translation enhancements through linguistic levels: A survey. ACM Comput. Surv. 46, 3, Article 42 (January 2014), 28 pages.
DOI: http://dx.doi.org/10.1145/2518130

1. INTRODUCTION

Machine Translation (MT) can be considered a highly interdisciplinary and multidisciplinary field because it is approached from the point of view of human translators, engineers, computer scientists, mathematicians and linguists. Nowadays, the cooperation and interaction between them are leading to interesting research outputs. Data-driven MT, such as Statistical Machine Translation (SMT), is prevalent within the MT academic research community and translation results obtained using these approaches have now reached a level of quality that make them useful for some particular applications. Given a parallel text at the sentence level, SMT uses probabilistic models to learn translations [Brown et al. 1993]. Given a source string, the goal is to choose the string with the highest probability among all possible target strings. Original word-based models have been replaced by phrase-based models [Zens et al. 2002; Koehn et al. 2003], which are directly estimated from aligned bilingual corpora by considering relative frequencies. Recent systems implement a general Maximum Likelihood Estimation (MLE)

This work is supported by the Seventh Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951).

Author's addresses: M. R. Costa-jussà, Human Language Technology Group, Institute for Infocomm Research, Singapore; email: vismrc@i2r.a-star.edu.sg; M. Farrús, N-RAS Center, Natural Language Processing Group, Universitat Pompeu Fabra, Barcelona; email: mireia.farrus@upf.edu.

DOI: http://dx.doi.org/10.1145/2518130

approach [Berger et al. 1996] in which a log-linear combination of multiple feature functions is used [Och and Ney 2002].

Another relevant MT paradigm is the rule-based (RBMT) [Forcada et al. 2009], which applies a set of linguistic rules in three different phases: analysis, transfer, and generation. Analysis and generation may be performed at different deeping linguistic levels from morphology, syntax up to semantics [Charoenpornsawat et al. 2002]. In the extreme, no transfer stage is needed when an interlingua language is used for representing source and target languages.

Many MT systems currently in use in industry are based on rules and are still actively investigated. Nowadays, the boundaries between rule-based and statistical MT approaches have narrowed and some approaches have been already proposed for constructing hybrid MT systems [España-Bonet et al. 2011; Thrumair 2009; Eisele et al. 2008].

The baseline Phrase-Based Statistical Machine Translation (PBSMT) approach faces translation in a holistic manner, and it makes little linguistic analysis of the language involved in the translation differently from RBMT. However, there are many studies that focus on how to enhance the PBSMT core system at the different linguistic levels. The growing number of studies and literature reflects the importance of getting linguistics involved in PBSMT. Moreover, the improvement achieved when going through linguistic information becomes evident [Costa-jussà et al. 2013].

According to the Linguistic Society of America [LSA 2013], linguistics is the scientific study of language. Depending on the linguistic properties of human language that are being analyzed, linguistics can be divided into several levels or subfields. These aspects or properties include sounds (phonetics, phonology), words (morphology), sentences (syntax), and meaning (semantics). It can also involve looking at how people use language in context (pragmatics, discourse analysis), or how to model aspects of language (computational linguistics), among others [LSA 2013]. Generally speaking, the following linguistic levels could be considered as the most prominent ones, according to their object of study:

—*Phonetics:* the study of the physical properties of human speech sounds.
—*Phonology:* the study of the sound system of a specific language or across languages.
—*Morphology:* the study of the internal structure of words and how they can be modified.
—*Lexis:* the study of the vocabulary of a particular language and their properties as the main units of language.
—*Syntax:* the study of the rules that govern the structure of grammatical sentences.
—*Semantics:* the study of the vocabulary meaning.
—*Pragmatics:* the study of how utterances are used in communicative acts, and the role played by context and non-linguistic knowledge in the transmission of meaning.
—*Discourse analysis:* the study of language in spoken, written, or signed texts.

The list of linguistic levels or subfields is not universal or unique. Some subfields can overlap considerable, or they can be combined leading to new subfields, or they can just be applied to other aspects of life or to other disciplines, leading also to new subfields. Subfields such as historical linguistics, sociolinguistics, dialectology, language acquisition, psycholinguistics, experimental linguistics, anthropological linguistics. and applied linguistics are also acknowledged by the Linguistic Society of America, but they are not relevant for the focus of this article. Regardless of any particular linguist's position or level classification, each area has core concepts that motivate significant studies and research.

The current article provides a deep analysis of PBSMT through the different linguistic levels, including orthography, morphology, lexis, syntax, and semantics. Orthography has been addressed by automatic correction of the parallel corpus used in translation or the translation output. Morphology, which is especially difficult to address in a PBSMT system when translating into a richer morphological language, has been introduced from different perspectives, including morpheme segmentation. Most relevant lexical challenges in PBSMT include the translation of unknown words, which most of the time require the use of extra resources. Syntactic challenges are faced by introducing linguistic technologies such as shallow or dependency parsing, which is shown to lead to a better translation performance. Finally, semantic enhancements include, among others, the introduction of Word Sense Disambiguation (WSD) techniques into the PBSMT core approach. These techniques reduce the sparseness of the data alleviating the problems at this semantic level.

Other standard linguistic levels such as phonology and phonetics are not taken into account because we are focusing on the written language translation. Pragmatics is not discussed either because, to the best of our knowledge, although there are works using pragmatics in MT [Helmreich and Farwell 1998], there are no works using pragmatics in PBSMT approaches yet. Finally, discourse analysis in SMT literature is not exhaustive, probably for the dimensions of the object of study, which involves information conveyed by segments larger than a single clause. Therefore, it will not be considered as a main contribution to this article. However, it is worth mentioning the works of Foster et al. [2010], Hardmeier and Federico [2010], Lenagard and Koehn [2010] and Meyer et al. [2012], that focus on how SMT can take advantage of discourse analysis [Webber 2012]. More information about the last findings on discourse in SMT can be consulted in Hardmeier's survey [Hardmeier 2012].

This article is structured as follows: Section 2 shows a brief overview of how linguistics needs have influenced the development of MT over the last decades. Section 3 revises the PBSMT approach, which, among the different SMT approaches, is the most popular one. Section 4 is the core of the article, in which a literature review of PBSMT using linguistic approaches is presented. Section 5 shows how linguistics has also been integrated in the SMT evaluation task; and, finally, conclusions are presented in Section 6.

2. LINGUISTICS IN STATISTICAL MACHINE TRANSLATION

> Modeling the mechanism of natural communication requires a description of language data which is empirically complete for all components of this theory of language, i.e., the lexicon, the morphology, the syntax, and the semantics, as well as the pragmatics and the representation of the internal context. [Hausser 2001].

In this statement, Hausser emphasized the importance of involving the different linguistic levels when dealing with the processing of natural communication. As part of natural language processing, SMT should not be exempt of it. However, this was not so at its beginnings. This section, and neither this article, is not the place to start with an extensive overview of the history of MT. However, it is important to briefly outline most items in the development of MT, in order to understand how linguistics came up and in which form.

It seems that the first attempts in creating mechanical dictionaries date back to the 17th century, although the first concrete proposal were not made until the 20th century in patents issued in 1933 by G. Artsrouni and P. Smirnov Troyanskii [Hutchins 1995]. Two decades later, the Georgetown-IBM experiment was held in New York, being the first public demonstration of an MT system. At that time, developments in linguistics such as generative linguistics and transformation grammar were proposed to improve the quality of the translations [Hutchins 2005]. The well-known ALPAC report in 1966 predicted no future for MT, so that research in the US was almost completely abandoned, and continued only in Canada, France, and Germany. Systran, Logos, and Meteo were practically the only translation systems developed in the 70s.

During the following decade, the research community made a step forward in terms of linguistic knowledge applied to MT. The main research relied on translation through some variety of intermediary linguistic representation involving morphological, syntactic, and semantic analysis. In the late 80s, novel statistical-based methods appeared, but the lack of syntactic and semantic rules was acknowledged [Hutchins 2005]. Then, the need of taking linguistic features into account when developing SMT systems became evident.

It is in the late 90s when linguistics really appears in the PBSMT paradigm. Syntax was introduced in 1997 in the work made by Wu [1997], who introduced alignment and segmentation tasks (among others) in *tree-to-tree* models. Soon after, lexis was introduced by the hand of Knight and Graehl [1998], who used transliteration to translate unknown name entities. Semantic approaches arised in García Varea et al. [2001] to enhance WSD by means of a Maximum Entropy approach in order to integrate contextual dependencies of both source and target sides. In 2003, morphological techniques appeared in PBSMT in the form of POS in the work of Ueffing and Ney [2003], and Koehn and Hoang [2007] introduced the factored-based translation inspired in the factored-based language models from Bilmes and Kirchhoff [2003]. Finally, the concept of *cognate* in the transliteration approach introduced by Kondrak et al. [2003] and Virga and Khundanpur [2003] in the orthographical field completed the list of linguistic levels used to overcome some of the problems in the PBSMT approach, which are, in turn, the focus of this article.

All the works cited in the previous paragraph are further explained in detail in Section 4, together with other related approaches, and classified into the five linguistic levels, which are the basis of this article: orthography, morphology, lexis, syntax, and semantics.

3. STATISTICAL MACHINE TRANSLATION: THE PHRASE-BASED APPROACH

There are several strategies that can be followed when translating between a pair of languages in SMT: phrase-based [Koehn et al. 2003], alignment templates [Och and Ney 2004], N-gram-based [Mariño et al. 2006], factored-based [Koehn and Hoang 2007], syntax-based [Yamada and Knight 2002] or hierarchical [Chiang 2007]. As follows, we briefly describe the phrase-based approach, which is the most popular SMT approach, and it has been described previously in other works [Costa-Jussà 2012]. Other approaches remain out-of-scope of this article.

Given a source sentence s, the SMT system chooses the target sentence $\hat{t}$ with the highest probability among all possible target sentence t, which is commonly known as the noisy channel approach to SMT [Brown et al. 1993].

The probability decomposition based on the Bayes' theorem allows to model independently target language and translation. The given source sentence s is segmented into sequences of one or more words, then each source segment is translated and the target sentence is composed from these segment translations. On the one hand, the translation model weights how likely words in the target language are translation of words in the source language; the language model, on the other hand, measures the fluency of hypothesis t. The search process is represented as a maximization operation of the product of both models.

The translation model in the phrase-based approach [Koehn et al. 2003] is composed of phrases. A phrase is a pair of m source words and n target words extracted from a parallel sentence that belongs to a bilingual corpus. The parallel sentences of the training corpus have previously been aligned at the word level [Brown et al. 1993]. Then, given a parallel sentence aligned at the word level, phrases are extracted as sequences of words consecutive in both source and target sides and consistent with the word alignment. A phrase is consistent with the word alignment if no word inside the phrase is aligned with any word outside the phrase. Finally, phrase translation probabilities are estimated as relative frequencies [Zens et al. 2002].

The language model assigns a probability to each target sentence. Standard language models are computed following the n-gram strategy, which considers sequences of n words. In order to compute the probability of an n-gram, it is assumed that the probability of observing the ith word in the context history of the preceding i-1 words can be approximated by the probability of observing it in the shortened context history of the preceding n-1 words. In addition, n-gram probabilities are computed using more complex techniques than counting known as smoothing techniques [Kneser and Ney 1995; Chen and Goodman 1996].

The noisy channel approach evolved into the log-linear model [Och and Ney 2002], which allows using several models or so-called features and to weight them independently. This approach should be interpreted as a maximum-entropy framework. The most common additional features that are used in this maximum-entropy framework (in addition to the standard translation and language model) are the lexical models, the word bonus and the reordering model. The lexical models are particularly useful in cases where the translation model may be sparse. For example, for phrases that may have appeared few times the translation model probability may not be well estimated. Then, the lexical models provide a probability among words [Brown et al. 1993] and they can be computed in both directions source-to-target and target-to-source. The word bonus is used to compensate the language model, which benefits shorter outputs. The reordering model is used to provide reordering between phrases, and there have been many proposed techniques for this [Costa-Jussà and Fonollosa 2009]. One of the most popular ones, for example, is the lexicalized reordering model [Tillman 2004], which classifies phrases by the movement they made relative to the previous used phrase. This movement can be either monotone, swapped, or discontinuous (MSD). Therefore, for each phrase, the model learns how likely it is followed by the previous phrase (monotone), swapped with it (swap) or not connected at all (discontinuous).

The different features or models are optimized in the decoder following the minimum error rate procedure described in Och [2003]. This algorithm searches for weights minimizing a given error measure, and it enables the weights to be optimized so that the decoder produces the best translations (according to some automatic metric and one or more references) on a development set of parallel sentences.

4. ANALYSING PBSMT THROUGH LINGUISTIC LEVELS

One of the main advantages of PBSMT over other kind of MT approaches is that it does not necessarily require linguistic knowledge. However, in practice, many works in the literature have shown that this type of knowledge can improve PBSMT systems. This section, which includes five subsections, shows the integration of different levels of linguistic knowledge (i.e., orthography, lexis, morphology, syntax, and semantics) into standard PBSMT systems. Therefore, most approaches mentioned here are directly applied to enhance the PBSMT, others are mentioned because they could be easily adapted to a PBSMT system and, finally, in Section 4.4, we show approaches beyond the PBSMT because the introduction of syntax knowlege in SMT is mostly done within the well-known syntax-based SMT approaches [Venugopal and Zollmann 2009]. Table I shows a summary of the PBSMT challenges through the linguistic levels, together with the main related works overviewed in this article.

Table I. Linguistic Challenges and Main Related Works

Linguistic level	Challenge	Main related works
Orthography	Spelling	Bertoldi et al. [2010], Farrús et al. [2011]
	Truecasing/Capitalization	Lita et al. [2003], Wang et al. [2006]
	Normalization	Riesa et al. [2006], Aw et al. [2006], Diab et al. [2007], Kobus et al. [2008]
	Tokenization	Farrús et al. [2011], El Kholy and Habash [2012]
	Transliteration	Boas [2002], Virga and Khudanpur [2003], Kondrak et al. [2003], Zhang et al. [2004], Kondrak [2005], Mulloni and Pekar [2006], Kumaran and Kellner [2007], Mitkov et al. [2007], Istvan and Shoichi [2009], Nakov and Ng [2009]
Morphology	Inflections	Brants [2000], Ueffing and Ney [2003], Creutz and Lagus [2005], Minkov et al. [2007], Koehn and Hoang [2007], Virpioja et al. [2007], Avramidis and Koehn [2008], de Gispert et al. [2009] El-Kahlout and Oflazer [2010], Bojar and Tamchyna [2011], Green and DeNero [2012], Formiga et al. [2012], Rosa et al. [2012]
Lexis	Unknown words	Knight and Graehl [1998], Al-Onaizan and Knight [2002], Koehn and Knight [2003], Fung and Cheung [2004], Shao and Ng [2004], Langlais and Patry [2007], Mirkin et al. [2009], Marton et al. [2009], Li et al. [2010], Huang et al. [2011], Zhang et al. [2012]
	Spurious words	Fraser and Marcu [2007], Li and Yarowsky [2008], Menezes and Quirk [2008]
Syntax	Word reordering	Wu [1997], Alshawi et al. [2000], Menezes and Richardson [2001], Yamada and Knight [2002], Aue et al. [2004], Galley et al. [2004], Ringger et al. [2004], Xia and McCord [2004], Chiang [2005], Collins et al. [2005], Ding and Palmer [2005], Quirk et al. [2005], Simard et al. [2005], Zhang and Gildea [2005], Galley et al. [2006], Liu et al. [2006], Huang et al. [2006], Langlais and Gotti [2006], Smith and Eisner [2006], Turian et al. [2006], Birch et al. [2007], Li et al. [2007], Zhang et al. [2007], Wang et al. [2007], Cowan [2008], Elming [2008], Graehl et al. [2008], Li and Yarowsky [2008], Badr et al. [2009], Genzel [2010], Shen et al. [2010], Khalilov and Fonollosa [2011], Bach [2012], Germann [2012]
Semantics	Sense disambiguation	García-Varea et al. [2001], Chiang [2005], Bangalore et al. [2007], Carpuat and Wu [2007], Chan et al. [2007], Carpuat and Wu [2008], Shen et al. [2009], Wu and Fung [2009], España-Bonet et al. [2009], Haque [2011], Banchs and Costa-jussà [2011], Banarescu et al. [2013]

4.1. Orthography

Orthography refers to the correct way of using a specific writing system to write a language. One of the first handicaps a PBSMT must deal with is the lack of orthographical consistency. Translating between languages written in different alphabets, facing with orthographical errors, or special writing registers are some of the challenges that arise be found in the translation task.

This section presents a brief overview of these types of problems that can be found in the recent literature, and the methods adopted in order to solve them.

4.1.1. Spelling Mistakes and Typographical Errors. A spelling mistake or a typographical error, even minor, will convert an existent word in the training corpus into an out-of-vocabulary word. Therefore, it is one of the issues to be addressed regarding orthographic aspects.

The methodology used in orthographic correction depends highly on the source and target languages, as well as the pair of languages involved. As an example of orthographic correction, Farrús et al. [2011] propose some solutions to overcome the orthographic errors in the Catalan-Spanish language pair, such as the incorrect use of the dot in the geminated *l*, the apostrophe, and the coordinating conjunctions *y* and *o*. The proposed solutions included a preprocessing based on text edition and the use of grammatical information. The geminated *l*, for instance, was corrected before translation by normalizing the writing of the middle dot. Other cases, such as the obligation *tener que* (*to have to*) and the conjunctions *y* and *o* (*and* and *or*), were corrected through post-processing after the translation. On the other hand, grammatical categories were used either in pre- or postprocessing rules (in order to solve problems with apostrophes, clitics, capital letters at the beginning of the sentences, relative pronouns, and semantic disambiguation) or in the translation model (for semantic disambiguation and lack of gender concordance).

Another paper worth citing for spelling correction is Bertoldi et al. [2010], which analyzes the impact of misspelled words in PBSMT. The authors propose an extension of the translation engine for handling misspellings, decoding a word-based confusion network representing spelling variations of the input text.

4.1.2. Truecasing and Capitalization. It is quite common in PBSMT to lowercase all training and testing data in order to avoid orthographic mismatchings. *Truecasing* is an alternative approach which aims at lowercasing only the words at the beginning of their sentence to their most frequent form. The work of Lita et al. [2003], for instance, discusses the truecasing process with an HMM. In this task, both a pre-processing step and a post-processing steps are required, in order to normalize the case and to further generate the proper surface forms. Prior to these processes, a truecasing model must be trained, which consists of a list of words together with the frequency of their different forms.

When PBSMT systems are trained on lowercased data, the case information needs to be recovered in a postprocessing step. This task is known as *recasing* or *capitalization*, and to this end, some systems such as Moses[1] provide simple tools to recase data. On the other hand, Wang et al. [2006] present a probabilistic bilingual capitalization model for capitalizing MT outputs using conditional random fields [Lafferty et al. 2001].

4.1.3. Normalization. Very often, the input text is generally correct, with no important spelling mistakes and typographical errors. However, some words can be usually written in different ways, leading to orthographic differences with respect to the trained corpus. In this case, an orthographic normalization is required, since it helps to improve

[1] http://www.statmt.org/moses/.

the translation quality because of the sparsity reduction they contribute, decreasing the number of out-of-vocabulary words.

One of the main linguistic issues that requires orthographic normalization is the Arabic *diacritization*. Diacritics in Arabic are optional orthographic symbols used to represent short vowels, and it is highly used in Arabic texts, although it depends partially on genre and domain. The impact of Arabic diacritization can be observed in Diab et al. [2007], in which several diacritization schemes ranging from full to partial diacritization are defined. It can be observed that the PBSMT performance is positively correlated with the increase in the number of tokens correctly affected by a diacritization scheme.

Another normalization method is the *contextual orthographic normalization*, presented by Riesa et al. [2006] for English-Iraqi Arabic speech-to-speech SMT system. Spelling errors and inconsistencies are very common in both languages, due to the lack of standard orthography and transliteration. On the English side, for instance, *Qoran*, *Qor'an,* and *Koran* are three different transliterations for the same proper name. Applying a global set of character-based normalization rules to a given text has the disadvantage of introducing many potential ambiguities in speech translation, since some of the characters eliminated or changed due to normalization generally carry important acoustic information for the posterior speech synthesis. In order to avoid this, the existence of shared semantics among words is decided by means of contextual analysis. Contextual orthographic normalization requires little linguistic knowledge and it can be easily adapted to other languages in which spelling or diacritical inconsistencies are common.

On the other hand, the language used in SMS, email, chats, and so on are far from the norm of the language. Several studies propose possible approaches to their automatic normalization [Kobus et al. 2008; Aw et al. 2006].

4.1.4. Tokenization and Detokenization. Tokenization is the process of splitting a stream of text up into appropriate elements or *tokens*. Tokenization is also about dealing with nonalpha characters like hyphens, apostrophes, punctuation, numbers, and others (phone numbers, URLs, emails, football scores, units, etc.). The main objective is to facilitate the input for further processing a text, which becomes essential in the MT task. Detokenization is thus the inverse process that takes place at the end of the translation task.

As normalization, tokenization also reduces sparsity and decreases the number of out-of-vocabulary words. It is not an orthographical correction method itself, but a morphological technique specially useful when dealing with rich morphological languages such as Arabic [El Kholy and Habash 2012]. However, detokenization is a complex process requiring many rules and exceptions. A good prior normalization is necessary in order to avoid problems derived from tokenization and detokenization. An example is found in Farrús et al. [2011], where there exists an incorrect use of apostrophes in languages such as Catalan and French, or spare blanks.

4.1.5. Transliteration. One of the problems that MT needs to deal with is the conversion of text strings from one orthography to another, while preserving the phonetics of the strings in both languages [Kumaran and Kellner 2007]. This task is known as *transliteration* and needs to be addressed, since most proper names are out-of-vocabulary words that need to be transliterated [Knight and Graehl 1998].

Cognates are words in different languages that are similar in their orthographic or phonetic form and are possible translations of each other, so that they are potential terms to be transliterated. Examples of cognates are *senhor* (Portuguese) vs. *señor* (Spanish), *apple* (English) vs. *Apfel* (German), or *vuit* (Catalan) vs. *huit* (French). Cognates can include names, numbers and punctuation, and they are defined by linguists as words derived from a common root. However, computational linguists tend to ignore the origin, and define cognates just as words with similar orthography that are mutual translations from each other [Nakov and Ng 2009]. It has been demonstrated that the use of cognates can improve SMT models [Kondrak et al. 2003; Mitkov et al. 2007; Kondrak 2005], and the out-of-vocabulary word problem. Therefore, much effort has been placed into detecting them in order to build bilingual dictionaries automatically [Boas 2002; Mulloni and Pekar 2006; Istvan and Shoichi 2009].

Although research literature on machine transliteration is not vast, numerous works can be found. In Kumaran and Kellner [2007], the transliteration task between a variety of different languages is addressed in a language-independent manner by using a statistical learning framework. Zhang et al. [2004] also present a novel framework for machine transliteration/back-transliteration that allows to perform direct orthographical mapping between source and target languages through an n-gram transliteration model. Nakov and Ng [2009] introduce a novel language-independent approach for improving PBSMT for resource-poor languages by exploiting their similarity to resource-rich ones. A resource-poor language X1 is translated into a resource-rich language Y, using parallel sentences between X2 and Y, being X2 a resource-rich language very closely related to X1. The method relies on the existence of a large number of cognates between X1 and X2, which often exhibit minor spelling variations, easy to learn automatically. Another work carried out by Virga and Khudanpur [2003] uses Chinese orthography to present a name transliteration procedure based on SMT techniques, for cross-lingual information retrieval purposes. The phonemic representations of English names are transliterated to a sequence of initials and finals. Then, another SMT model is used to map the obtained initial/final sequence to Chinese characters.

4.2. Morphology

Morphology refers to identification, analysis and description of the word internal structure. The challenges raised when translating from or into richer morphology languages are well known and are being continuously studied in the context of PBSMT. Morphology is the study of the structure of a set of given language morphemes, such as stems or affixes [Karageorgakis et al. 2005]. Although the most important morpheme is the stem, in this article we will deal with morphemes other than the stem. These morphemes provide syntactic information about tense, count, case, gender, function, and so on (e.g., the word *older* consists of the stem *old* and the comparative affix *-er*). Even irregular forms can be represented using these morphemes, although they usually are not represented by the typical forms. So, for instance, the word *better* consist of the morpheme *good* and a comparative morpheme.

Morphologically rich languages have many different surface forms, even though the stem of a word may be the same. This leads to rapid vocabulary growth, as various prefixes and suffixes can combine with stems in a large number of possible combinations and worse language model probability estimation because of more singletons (forms occurring just once in the data), and a lower number of occurrences over all distinct words. The sparsity due to morphology can be reduced by incorporating morphological information into the PBSMT system. The three most common solutions go through: (1) a preprocessing of the data so that the input language more closely resembles the output language; (2) the use of additional feature functions that introduce morphological information; and (3) a postprocessing of the output to add proper inflections. The following subsections refer to the research work from each type of solution.

4.2.1. Preprocessing the Data. The idea here is to preprocess the data so that the input language more closely resembles the output language, by means of either enriched input models or segmented translation.

Enriched input models tend to focus on the problem of going from a less inflected language into a higher inflected one. This type of approaches try to solve the challenge that word forms often do not contain enough information for producing the correct full form in the target language. Ueffing and Ney [2003] use POS (Part-Of-Speech) tags as additional source knowledge and enrich the English verbs such that they contain more information relevant for selecting the correct inflected form in the target language. The lexicon model is then trained using the maximum entropy approach, taking the verbs as additional features. Avramidis and Koehn [2008] focus on two linguistic phenomena, which produce common errors on the output, that is, noun cases and verb persons. Their algorithm uses heuristic syntax-based rules on the statistically generated syntax tree of each sentence, in order to address the missing information, which is consequently tagged in by means of word factors. This information is proven to improve the outcome of a factored PBSMT model, by reducing the grammatical agreement errors in the generated sentences.

On the other hand, regarding segmented translation, El-Kahlout and Oflazer [2010] experiment with a morphemic representation of the parallel texts and align the sentences at the morpheme level. Additionally, in order to help with word alignment, they experiment with local word reordering to bring English prepositional phrases and auxiliary verb complexes in line with the morpheme order of the corresponding Turkish order. The decoder produces stems and morpheme sequences, which are then selectively concatenated into surface words. However, they only show improvements when performing a simple grouping of stems and morphemes, which is performed by extracting frequently occurring n-grams. This grouping is complemented with a selective morpheme concatenation that only allows to combine those morphemes and stems that form a valid Turkish word form as checked by a morphological analyzer. Similarly, Virpioja et al. [2007] use the Morfessor algorithm [Creutz and Lagus 2005] to find statistical morpheme-like units that can be used to reduce the size of the lexicon and improve the ability to generalize. Translation and language models are trained directly on morphemes instead of words. The approach is tested over three Nordic languages (Danish, Finnish, and Swedish). Although, the proposed solution does not improve in terms of BLEU [Papineni et al. 2002], it has clear benefits, as morphologically well motivated structures (phrases) are learned, and the proportion of untranslated words is significantly reduced. A more recent publication with the use of the Morfessor algorithm [de Gispert et al. 2009] shows better BLEU by using Minimum Bayes Risk (MBR) combination.

4.2.2. Additional Algorithms or Feature Functions. Several works introduce additional feature functions to improve morphology in PBSMT. Green and DeNero [2012], for instance, use a class-based agreement model for generating accurately inflected translations. Agreement is found by scoring a sequence of fine-grained morpho-syntactic classes that are predicted during decoding for each translation hypothesis.

Other approaches use additional language models by means of the factored-based translation [Koehn and Hoang 2007]. Inspired in the factored-based language models [Bilmes and Kirchhoff 2003], the factored-based approach is an extension of the phrase-based approach presented in Section 3. It adds additional annotation at the word level. A word in this framework is not anymore only a token, but a vector of factors that represent different levels of annotation such as lemmas and POS.

As explained in Koehn et al. [2007], the translation of factored representations of input words into the factored representations of output words is broken up into a sequence of mapping steps that either translate input factors into output factors, or generate additional output factors from existing output factors. Factored translation models follow closely the statistical modeling approach of phrase-based models (in fact, phrase-based models are a special case of factored models). The main difference lies in the preparation of the training data and the type of models learned from the data. Most experiments on factored translation models use the POS factors and the generation of POS for a specific corpus is usually done by statistical tools trained on specific corpus that have been manually tagged by expert annotators. The POS marks the tokens with their corresponding word type based on the token itself and the context of the token. A token can have multiple POS depending on the token and the context. A standard algorithm for POS taggers with reasonable accuracy (97%–98% tested in English) is the HMM-based tagger described by Brants [2000].

4.2.3. Postprocessing the Translation Output. Postprocessing the output of a PBSMT system allows to add on the proper inflections by means of morphology generation [Minkov et al. 2007; Bojar and Tamchyna 2011; Formiga et al. 2012; Rosa et al. 2012]. These approaches factor the problem of translation into two subproblems: predicting stems and predicting inflections. Minkov et al. [2007] use stems and inflection prediction done by means of Maximum Entropy Markov models. Similarly, Formiga et al. [2012] simplify the target language using stems. They build a PBSMT system, which considers morphology generation as an independent natural language processing task. They only focus on verbs and the morphology generation task is addressed as a multiclass classification problem which uses shallow and deep features. Their approach achieves better generalizations in out-of-domain data. Bojar and Tamchyna [2011] approach is based on training a factored PBSMT system in the reverse direction and translating a large monolingual corpus using this system. This generates a new parallel data that is added to retrain the system. To learn new target forms, the monolingual target corpus is used both with full word forms and with lemmas. Finally, Rosa et al. [2012] present a system for automatic rule-based postprocessing of English-to-Czech MT outputs using a parser. The set of rules fixes structure, agreement, translation, and other minor issues such as capitalization in the target sentence.

4.3. Lexis

Lexis refers to the set of words and phrases of a particular language. In recent years, the compilation of language databases using real samples has allowed researchers to study the language lexicon and how it is composed. The statistical research methods show how words interact. However, there are several challenges in MT coming up at this level due to using this statistical methods. In this section, we report them and we show the state-of-the-art solutions.

4.3.1. Unknown Words. In the area of MT, almost all of the literature focus on finding the correct translation of unknown words either with external resources and/or lexical rules. Other methods using morphology have been already shown in Section 4.2.

Early approaches in this issue like Knight and Graehl [1998] and Al-Onaizan and Knight [2002] make use of transliteration and web mining techniques with external data to translate unknown Name Entities (NEs). Koehn and Knight [2003] translate the compound unknown words by splitting them into in-vocabulary words. Following studies carried out by Fung and Cheung [2004] and Shao and Ng [2004] adopt comparable corpora and web resources to extract translations for each unknown word. Later on, Langlais and Patry [2007] use analogical learning for translating unknown words; Mirkin et al. [2009] apply entailment rules and Marton et al. [2009], a paraphrase model to replace unknown words with in-vocabulary words using large set of additional bitexts or manually compiled synonym thesaurus WordNet. Li et al. [2010] address the problem of translating numeral and temporal expressions. They used manually created rules to recognize the numeral/temporal expressions in the training data and replaced them with a special symbol. Consequently, both of the translation rule extraction and reordering model training consider the special symbol. In the decoding time, if a numeral or temporal expression is found, it is substituted by the special symbol so that the surrounding words can be handled properly and finally the numeral/temporal expression is translated with the manually written rules. Huang et al. [2011] propose a sublexical translation method to translate Chinese abbreviations and compounds. More recently, Zhang et al. [2012] address the problem of the lexical selection and word reordering of the surrounding words caused by unknown words. They consider all kinds of unknown words and apply a distributional semantic model and a bidirectional language model to fulfill this task without any additional resources.

4.3.2. Spurious Words. These are words that do not have any counterpart in other languages. An MT system should be able to identify the spurious words of the source language and not translate them, as well as to generate the spurious words of the target language. By default, PBSMT systems allow a source language phrase to translate to nothing or to capture the source word deletion inside a phrase pair. Li and Yarowsky [2008] use a specific empty symbol on the target language side and any source word is allowed to translate into it. This symbol is invisible in every module of the decoder except in the translation model. That means that it is not counted when calculating language model score, word penalty, and any other feature values, and it is omitted in the final output of the decoder. It is only used to delete spurious source words and refine translation model scores accordingly.

Other approaches to deal with spurious words are introduced in the word alignment procedure [Fraser and Marcu 2007] or in other type of SMT systems different than the PBSMT [Menezes and Quirk 2008].

4.4. Syntax

Syntax refers to the principles and rules for constructing sentences in natural language. This term is popularly used to refer to the rules that determine the sentence structure of a particular language. Basic PBSMT systems do not include this type of information because phrases in these models are just sequences of words with no structure. One of the highest consequence derived from not using syntax information is the word reordering errors when translating into more fixed-order languages like English. In free word-order languages, reordering becomes less important and there are more errors in morphological agreement between syntactically dependent words. In any case, there are alternatives to the standard PBSMT systems that use statistical parsers to introduce syntax knowledge. This section overviews the most popular syntactic techniques that are used in the PBSMT systems, but unlike the other linguistic levels, the most important work here in SMT is done beyond the PBSMT. Syntax knowlege in SMT is covered by the syntax-based SMT approaches such as string-to-tree, tree-to-string, or tree-to-tree, shown in Figure 1, this section briefly reports these approaches. Related survey reports that we have taken into account here are provided by Razmara [2011] and Ahmed and Hannemann [2005].

4.4.1. Syntactic Techniques in PBSMT Systems. Syntax has failed to be introduced in the PBSMT systems in an approach where the phrases from the alignment were filtered to remove any phrases that do not correspond to a grammatical constituent [Koehn et al. 2003]. However, syntax knowledge has been successfully introduced in PBSMT systems specially to face reordering challenges. Most cases compute a prereordering, which can be either deterministic or nondeterministic. Deterministic prereorderings have been proposed by Xia and McCord [2004], Collins et al. [2005], Wang et al. [2007], Badr et al. [2009] and Genzel [2010], who use syntactic parsing and describe a set of syntactic reordering rules that exploit systematic differences between source and target word order. The resulting system is used as a preprocessor for both training

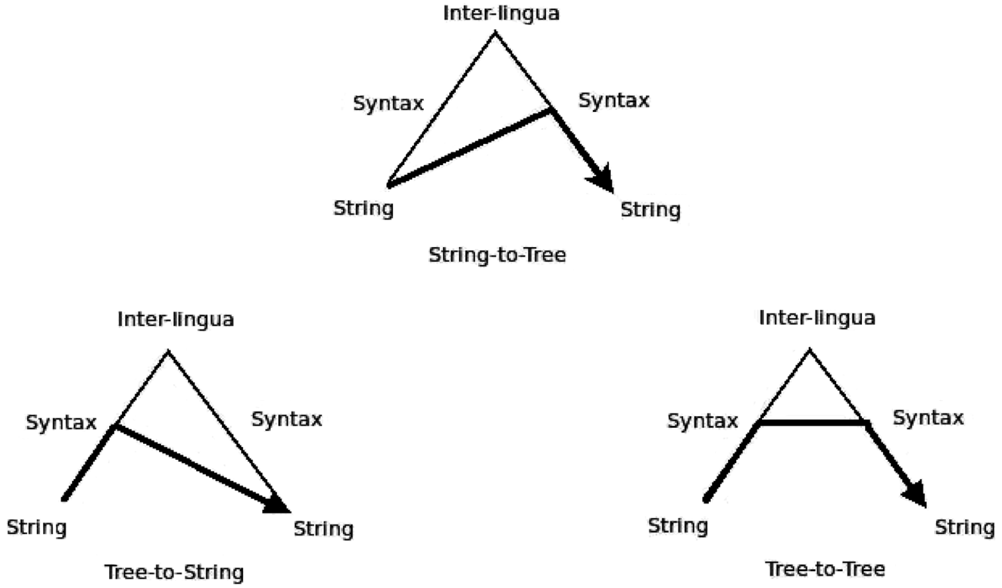

Fig. 1. Syntax-based SMT approaches.

and test sentences, transforming source sentences to be much closer to target in terms of their word order. Nondeterministic prereorderings, which basically offer different reordering schemes to the PBSMT decoder for it to take the decision, have been presented by other authors such as Li et al. [2007], Elming [2008], Khalilov and Fonollosa [2011] or Germann [2012]. Elming's [2008] proposes automatic reordering rule learning based on a rich set of linguistic information. Similarly, Khalilov and Fonollosa [2011] alleviate the word order challenge including morpho-syntactic and statistical information in the context of a pretranslation reordering framework aimed at capturing short- and long-distance word distortion dependencies. Recently, Germann [2012] presented a variant of PBSMT that uses source-side parsing and a constituent reordering model based on word alignments in the word-aligned training corpus to predict hierarchical block-wise reordering of the input. They build multiple possible translation orders in a source order lattice, which is then annotated with phrase-level translations to form a lattice of tokens in the target language. They propose various feature functions to evaluate paths through that lattice.

In a different way, Birch et al. [2007] use Combinatorial Categorical Grammar (CCG) supertags into a PBSMT system with an opportunity to access rich syntactic information at the word level. This approach is related to the factored models approach from Section 4.2.2.

4.4.2. String-to-Tree Models. These models leverage a monolingual parse tree at the target side from a source string. Yamada and Knight [2002] give a string-to-tree SMT approach which model is described in details how the target-tree is stochastically transformed to the source string. The intuition here is that these steps would model mapping from Subject Verb Object (SVO) languages to SOV ones. Decoding is modeled as parsing the source side to get the target tree. Galley et al. [2004] propose to learn a direct translation model that maps source strings to target trees, which uses rules that condition on a larger tree-fragment. In a further study, Galley et al. [2006] use

Expectation Maximization (EM) to learn rule probabilities using a variant of the generic tree-transducers learning framework.

Simard et al. [2005] propose a phrase-based model that allows for phrases with gaps. Training is done without limiting phrase extraction to contiguous ones. A more complete approach that allows for phrase gaps is the work by Chiang [2005], which gives a heuristic approach to learning Synchronous Context Free Grammars (SCFG) on top of the output of a phrase-based model, which is known as hierarchical SMT. This popular approach has gained many adepts and it has been followed and extended by many authors [Chiang 2007; Chiang et al. 2009; Hoang and Lopez 2009; Vilar 2011], and it can be seen as a generalization of the PBSMT approach that allows phrases with the ability to have sub-phrases and decoding is modeled as a parsing process. Shen et al. [2010] use a string-to-dependency algorithm, which employs a target dependency language model during decoding to exploit long distance word relations.

4.4.3. Tree-to-String Models. These models leverage a monolingual parse tree at the source side to a target string. Langlais and Gotti [2006] extend phrase-based models by concatenating together Tree-Phrases (TP), that is, associations between simple syntactic dependency treelets in a source language and their corresponding phrases in a target language where treelet can be defined to be an arbitrary connected subgraph of the dependency tree. The TP they use are syntactically informed and present the advantage of gathering source and target material whose words do not have to be adjacent. They parse the source side of the parallel corpus to produce a dependency-based parse trees, and then they align two strings. To extract those TPs, the source dependency trees were broken into treelets of depth one (head and its modifiers). The part of the target string that align with lexical items in this treelet is attached to form the TP pair. This TP pair can have gaps on both the source and target sides. TP probabilities are calculated using relative frequencies. A typical phrase-based decoder is then used in a left-right fashion by adding to the target string one phrase at a time. Liu et al. [2006] use a translation model based on tree-to-string alignment template. A source sentence is translated by using a parser to produce a source parse tree and then applying tree-to-string alignment templates to transform the tree into a target string. This tree-to-string alignment template is in charge of generating terminals and non-terminals and performing reordering at low and high levels. Huang et al. [2006] uses a log-linear framework allowing to rescore with other features including language models. Further work [Li and Yarowsky 2008] covers the idea of forest-based translation that allows to extract rules from a packed forest that compactly encodes exponentially many parses.

4.4.4. Tree-to-Tree Models. These models leverage a monolingual parse tree at both target and source sides. The source language input is parsed into a syntactic tree structure and the source language tree is mapped to a target language tree. The main advantage is that parsing the input generates valuable information about its meaning. In addition, the mapping from a source language tree to a target language tree helps preserve the meaning of the input and produce a grammatically correct output. A key disadvantage of this approach is that it requires a parser in both languages, which restricts the use of language pairs. Wu [1997] uses the formalism of Inversion Transduction Grammars (ITG) inducing alignment, segmentation tasks and other, whereas Zhang and Gildea [2005] give a lexicalized version of ITG. Alshawi et al. [2000] treat translation as a process of simultaneous induction of source and target dependency trees using head-transduction. Menezes and Richardson [2001] parses both source and target languages to obtain a Logical Form (LF), and translates source LFs using memorized aligned LF patterns to produce a target LF and it uses a separate sentence realization component [Ringger et al. 2004] in order to turn this into a target sentence. Aue et al. [2004] incorporated a LF-based language model into the system.

Quirk et al. [2005] align a parallel corpus, project the source dependency parse onto the target sentence, extract dependency treelet translation pairs, and train a tree-based ordering model. The word alignments are used to project the source dependency parses onto the target sentences. From this aligned parallel dependency corpus they extract a treelet translation model incorporating source and target treelet pairs. A unique feature is that they allow treelets with a wildcard root, effectively allowing mappings for siblings in the dependency tree. They also train a variety of statistical models on this aligned dependency tree corpus, including a channel model and an order model. In order to translate an input sentence, they parse the sentence, producing a dependency tree for that sentence. Then, they employ a decoder to find a combination and ordering of treelet translation pairs that cover the source tree and are optimal according to a set of models that are combined in a log-linear framework.

Ding and Palmer [2005] present an approach based on recursively splitting dependency trees. Turian et al. [2006] utilized a discriminative training approach utilizing regularized decision tree ensembles. Smith and Eisner [2006] also use dependency trees on both sides, but they allow for a more “sloppy” transfer rules that could capture a wider range of syntactic movements. Zhang et al. [2007] use tree-to-tree alignment between a source parse tree and a target parse tree. The model is formally a probabilistic synchronous tree-substitution grammar that is a collection of aligned elementary tree pairs with mapping probabilities (which are automatically learned from word-aligned bi-parsed parallel texts). This model supports multilevel global structure distortion of the tree typology and can fully utilize the source and target parse tree structure features. Graehl et al. [2008] basically address the training problem for probabilistic tree-to-tree transducers by giving a generic tree-transducer learning algorithm that utilizes an EM algorithm augmented with a modified inside-outside dynamic programming scheme to scale the E-step. Cowan et al. [2008] propose a method for the extraction of syntactic structures with alignment information from a parallel corpus of translations, and they make use of a discriminative, feature-based model for prediction of these target language syntactic structures. Recent works [Bach 2012] focus on the integration of dependency structures into MT components including decoder algorithm, reordering models, confidence measure, and sentence simplification.

4.5. Semantics

Semantics is the study of the meaning of words and phrases and the combination between them. This part of linguistics is not directly included in the PBSMT core algorithm, which means that semantic challenges such as homonymy/polysemy (i.e., the same word having different unrelated/related meanings depending on the context) or synonymy (i.e., different words having the same meaning) or semantic role labels are not specifically dealt with. Therefore, either they are learned directly from data, they are incorrectly translated, or they are not translated. One could discuss that the idea of probabilistic translation models is motivated by different word choices due to different senses of the input word. In practice, word context taken into account to translate a word may be insufficient when a word has multiple meanings.

Often we start with lexical semantics. To translate a word correctly, we need to know what it means. Word Sense Disambiguation (WSD) uses the input context to predict the ambiguous concepts. In the machine learning area, there is high quantity of literature dedicated to this issue, including popular evaluation campaigns such as the semantic evaluation series of workshops (SemEval) that focuses on the evaluation of semantic analysis systems, with the aim of comparing systems that can analyze diverse semantic phenomena in text.

Beyond lexical semantics, there are the semantic role labels, which study the meaning of a complex expression as a function of its parts. To translate a sentence correctly, we need to understand the objects and their relationships. Confusion of semantic roles causes translation errors that often result in serious misunderstandings of the essential meaning of the source utterances who did what to whom, for whom or what, how, where, when, and why. Recent works on this area show promising improvements as reviewed in Wu [2009].

This section focuses on how, recently, both lexical semantics and semantic role labelling have been introduced in statistical-based systems solve WSD by either using source or target context information. Most popular approaches make use of machine learning techniques or additional resources such as semantic parsers. Just recently, there is an approach that aims at full sentence semantics. Main research in this direction has started with the construction of an Abstract Meaning Representation Bank, which is a set of English sentences paired with simple, readable semantic representations [Banarescu et al. 2013].

4.5.1. Lexical Semantics. Early approaches on this issue are related to integrating contextual dependencies while training a discriminative word alignment like García Varea et al. [2001] who use a Maximum Entropy approach to integrate contextual dependencies of both source and target sides of the statistical alignment model.

More recent approaches for PBSMT systems, like Carpuat and Wu [2007], integrate contextual dependencies directly in translation and design a context-dependent lexicon that is matched to a given PBSMT model. The key idea is the fact that phrases should be context-dependent, taking into account the dynamic full sequence context as registered by richer features (including bag-of-words, local collocations, position-specific local POS and basic dependency features). Further extensions of this work can be found in Carpuat and Wu [2008],where the authors propose dynamically built context-dependent phrases instead of conventional static phrases, which ignore all contextual information. This model succeeds by making the following three adaptations, as mentioned in Wu [2009]:

(1) The WSD model is trained to predict observable senses that are the direct lexical translations of the target lexeme being disambiguated.
(2) WSD is redefined to move beyond the particular case of single-token and to generalize to multitoken.
(3) The WSD is fully integrated into the runtime decoding.

Haque [2011] proposes the use of a range of contextual features, including lexical features of neighboring words, POS tags, supertags, dependency information, that is, different syntactic and lexical features for incorporating information about the neighbouring words. Similarly, España-Bonet et al. [2009] train local classifiers, using linguistic and context information, to translate a phrase. The contextual features in this work are defined by taking into account words of the immediate context, n-grams, POS, lemmas, chunk label, and bag-of-words.

For other MT systems, there are works like Bangalore et al. [2007] that integrate Maximum Entropy models considering n-gram features from the source sentence. Chan et al. [2007] integrate WSD into a hierarchical phrase-based system, HIERO [Chiang 2005] by introducing two additional features into the MT model, which operate on the existing rules of the grammar without introducing competing rules. Also, in the same type of translation, Shen et al. [2009] use features with nonterminal labels and length distribution, source context, and a language model created from source-side grammatical dependency structures.

Banchs and Costa-jussà [2011] exploit similarity measures for computing the source context similarity between the input sentence to be translated and the original training material. The authors exploit both the standard vector-space model [Salton and McGill

1983] and latent semantic indexing [Landauer et al. 1998]. The objective of the proposed features is to account for the degree of similarity between a given input sentence and each individual sentence in the training dataset. The computed similarity values are used as an additional feature in the log-linear model combination approach to PBSMT. This model aims at favoring those translation units that were extracted from training sentences that are semantically related to the current input sentence being translated.

4.5.2. Semantic Role Labeling. The SRL should be useful in MT because they generally agree between the source and target languages and they guide the main structure of a sentence. A main approach in this area is from Wu and Fung [2009], which exploits semantic parsers, labels automatically the predicates and roles of the various semantic frames in a sentence and identify inconsistent semantic frame and role mappings between the input and the output sentences.

5. INTEGRATION OF LINGUISTICS INTO THE SMT EVALUATION TASK

One of the major needs in the MT field has been to find an appropriate system evaluation procedure to tune and test the quality of the output translations. During the last years, two very different ways of evaluating MT systems have appeared within the research community. On the one hand, there are a considerable number of automatic evaluation methods like bilingual evaluation understudy (BLEU [Papineni et al. 2002]), word error rate (WER [McCowan et al. 2004]), and translation error rate (TER [Snover et al. 2009]). METEOR [Lavie and Agarwal 2007], which is also becoming quite popular, is able to produce detailed word-to-word alignments between the system translation and the reference translation, which can help in the error analysis task. The main handicaps of these methods are that manual references cannot cover all possible translations.

On the other hand, human evaluators have been widely used to analyze the performance of the systems by means of their perception of the translation quality. These methods are based on a pairwise comparison of systems (e.g., Bojar et al. [2011]), where the annotator is asked to choose the best translation. Normally, given a translation output, a source sentence and a reference sentence, the evaluator is asked to score an output sentence between 1 (lowest score) and 5 (highest score) in two different evaluation metrics: adequacy and fluency.

However, few of the state-of-the-art automatic evaluation methods use the linguistic knowledge to evaluate SMT systems. Only some proposals regarding evaluation classification schemas can be found in the literature as alternatives to the aforementioned traditional methods. Nevertheless, at the same time the use of linguistic knowledge is growing in the different SMT approaches, so does the use linguistic knowledge in the evaluation task. Certainly, evaluation is an essential part in the translation task, and if the tendency of using hybrid approaches involving linguistic features in statistical systems has currently a renewed interest, the evaluation task must not be exempt of incorporating this linguistic knowledge.

Evaluation based on linguistic features is usually a list of categories to be analyzed in the translation output, in order to determine the correctness of the category taken into account. This type of evaluation has the advantage of being more informative, in the sense that we know, in the end, what types of errors are the most prominent in our system, so that we will be able to focus more on them by choosing the correct linguistic approach.

Normally, these categorization or error classifications can be *language-pairdependent* or *language-pair independent*. Language-pair–dependent classifications have the advantage of being more specific, and thus more reliable. Language-pair–independent classifications, instead, can lose concreteness but they are generalizable to any pair of languages. On the other hand, linguistic evaluation methods can also be classified

Table II. Linguistic Evaluation Methods in SMT Evaluation Task

AUTHOR	MAIN CATEGORIZATION	CHARACTERISTICS
Flanagan [1994]	spelling, not found word, accent, capitalization, inflection, article, pronoun, preposition, etc.	manual, language-pair dependent (English-French, English-German)
Vilar et al. [2006]	missing words, word order, incorrect words, unknown words, punctuation	manual, language-pair dependent (Spanish-English, English-Spanish, Chinese-English)
Popović et al. [2006]	syntactic differences (nouns & adjectives), Spanish inflections (verbs, adjectives, and nouns)	automatic, language-pair dependent (Spanish-English)
Giménez and Màrquez [2007]	lexis, shallow-syntaxis, syntaxis, shallow-semantics	automatic, language-pair independent, automatic metrics not limited to lexis
Popović and Ney [2009]	syntactic structure of the sentence	automatic, language-pair independent, based on BLEU, TER and METEOR metrics
Farrús et al. [2010]	orthography, morphology, lexis, semantics, syntaxis	manual, language-pair dependent (Catalan- Spanish), generalizable to any lang. pair
Birch and Osborne [2010]	lexical quality reordering quality	automatic, language-pair independent, correlated with human judgements
Lo and Wu [2011]	semantic role fillers	semiautomatic, correlated with human adequacy judgements
Popović and Ney [2011]	inflectional, word order, missing words, extra words, incorrect lexical choices	automatic, language-pair independent, based on WER and PER, correlated with human judgements

as manual or automatic. Again, manual (or human) methods gain reliability and they are time-consuming, whereas automatic methods are much faster and less concrete. This section presents a review of the main linguistic evaluation methods found in literature, classified into language-pair dependent and language-pair–independent. In Table II, all these methods are chronologically presented, together with their main characteristics.

5.1. Language-Pair–Dependent Lnguistic Evaluation Methods

5.1.1. Manual Classifications. One of the precedent error classifications in SMT is probably the one introduced by Flanagan for MT, [1994]. In the Flanagan classification, the errors are assigned to different categories, in order to provide a descriptive framework that can reveal relationships between errors, and to help the evaluator to map the extent of the effect in chains of errors. These categories are language-pair dependent, and they are set for the English-French and English-German pairs, including spelling, not found word, accent, capitalization, elision, verb inflection, noun inflection, other inflection, rearrangement, category, pronoun, article, preposition, negative, conjunction, agreement, clause boundary, word selection, expression, relative pronoun, case, and punctuation.

Vilar et al. [2006] propose a five-category main schema including missing words, word order, incorrect words, unknown words, and punctuation. At the same time, this big classification includes subtypes of errors, for example, missing words are classified into *content* and *filler* words, word order is seen at both phrase and word levels, and so on. This error classification was tested in the first evaluation campaign of the European TC-STAR project, from which it could be concluded that, although the big classification could be applied *a priori* to any pair of languages, the most important class of errors was language-pair dependentfor example, the verb tense generation for translation from English into Spanish, or the word order for translation from Chinese into English.

As far as we know, the unique linguistic evaluation method existing for the Catalan-Spanish pair is the one found in Farrús et al. [2010]. This method is based on the assumption that all errors can be classified into one of the following linguistic levels: orthographic, morphological, lexical, syntactic, and semantic. At the same time, every single level has a list of language-pair–dependent errors that can be found in the translation output. However, one of the advantages of this evaluation method is that the main categories (orthography, morphology, lexis, syntax, and semantics) can be seen as language-pair independent so that the analysis at the main category levels can be easily compared between different languages. Moreover, the results found in [Farrús et al. 2010] show that, after a manual annotation of the output errors, some linguistic error levels can be associated more closely to a human perceptual evaluation than others. A further research in Farrús et al. [2012] shows that the semantic level has a closer correlation with both human perceptual evaluation and automatic metrics than the other linguistic levels.

The error typologies proposed by Flanagan [1994], Vilar et al. [2006], and Farrús et al. [2010] have been implemented in the BLAST (the BiLingual Annotator/Annotation/Analysis Support Tool) system [Stymne 2011], an open- source tool for error analysis and human annotations of bilingual material extracted from MT output.

5.1.2. Automatic Classifications. Shortly after the classification made by Vilar et al. [2006], the importance of using linguistic information was acknowledged by Popović et al. [2006], together with the need of automatizing the process, considering that a human error analysis and error classification was a very time consuming task. Therefore, the authors propose the use of morpho-syntactic information in combination with the automatic evaluation measures WER and PER in order to get more details about the translation errors. This morpho-syntactic information includes (a) syntactic differences between Spanish and English taking into account nouns and adjectives, and (b) Spanish inflections related mainly to verbs, adjectives, and nouns.

5.2. Language-Pair–Independent Linguistic Evaluation Methods

As far as we are concerned, language-pair–independent evaluation methods using linguistic knowledge have only been developed as an automatic task. The main motivation is based on the fact that traditional automatic metrics such as BLEU limit their scope to the lexical dimensions. In this sense, Giménez and Màrquez [2007] suggest to use new metrics that take into account linguistic features at more abstract levels, based on the assumption that lexical similarity is nor a sufficient neither a necessary condition so that two sentences convey the same meaning. The authors adapt a wide set of metrics for automatic MT evaluation at four linguistic levels: lexical, shallow-syntactic, syntactic, and shallow-semantic under different scenarios, showing that linguistic features at more abstract levels may provide more reliable system rankings.

Popović and Ney [2009] present a framework for automatic error analysis and categorization. In some of their previous works [Popović and Ney 2007], the basic idea is to identify erroneous words using algorithms for the calculation of WER and PER. The extracted error details are used in combination with several types of natural language knowledge, such as base forms, POS tags, and others. Here, the hypothesis is extended to BLEU, TER, and METEOR and oriented to the syntactic structure of the sentence. Although the new metric measures can be applied to any pair of languages, they are tested over the outputs of translation from Spanish, French, and German into English and vice versa. Results show a competitive performance with respect to the traditional BLEU, METEOR, and TER metrics, as well as a high correlation with human judgements.

Later, Popović and Ney [2011] proposed a new framework for automatic error analysis and classification using the algorithms for WER and PER. The analysis is focused on five main error categories: inflectional errors, errors due to wrong word order, missing words, extra words and incorrect lexical choices, and the contribution of various POS classes is taken into account. This error analysis was tested over Arabic-English, Chinese-English, Spanish-English and German-English outputs generated in the framework of the Newswire and Broadcast News, the GALE[2] project, the TC-STAR[3] project, and the Fourth Workshop on Statistical Machine Translation[4] (WMT'09), respectively. Again, a high correlation with human judgements was found.

The work of Birch and Osborne [2010] is based on the assumption that the traditional MT metrics do not adequately measure the reordering performance of translation systems. In this work, the authors present the LRscore metric to evaluate the lexical and reordering quality in SMT, which, apart from being language independent, it showed to be much more consistent with human judgements than BLEU. Finally, the semiautomatic metric MEANT introduced by Lo and Wu [2011], assesses translation utility by matching semantic role fillers. The scores produced correlate with human judgment.

Finally, there are recent approaches that use quality estimation as a quality indicator of translation outputs, and the main difference with machine translation evaluation is that they do not rely on reference translation and usually rely on machine learning methods together with linguistic features to provide quality scores [Felice and Specia 2012].

6. CONCLUSIONS

Research in the field of SMT is nowadays evolving into the concept of hybridization, though in two different—but clearly related—ways. On the one hand, hybrid systems are seen as a combination of statistical systems with existing rule-based systems. On the other hand, there is a growing interest in combining linguistic knowledge in all its forms (e.g., morphological, syntactic, and semantic) into the existing statistical systems.

The current article has presented an overview of how to overcome some of the problems encountered in SMT, especifically in PBSMT, through five linguistic levels: orthography, morphology, lexis, syntax, and semantics. As it can be concluded from the current state of the art, the performance of SMT systems can be clearly improved by using such linguistic knowledge. Nevertheless, the holistic SMT is still not able to cover correctly all the translation challenges that arise from the statistical systems. Alternatively, instead of being general, each extension to SMT tends to focus on one particular challenge to achieve the desired enhancement, and these particular approaches are usually focused on one of the linguistic levels mentioned earlier.

Additionally, linguistic knowledge has also been brought to the evaluation task, an essential part of the MT process. Several error typologies have been proposed, some of them directly based on automatic measures such as BLEU, TER, WER, and PER, and others based on more pure linguistic criteria. In any case, there is a clear tendency to use linguistic information in the evaluation task and to automatize the error categorization. The use of evaluation metrics that take into account several linguistic levels adds objectivity to the evaluation and it usually achieves a higher correlation with human judgements.

[2] GALE: Global Autonomous Language Exploitation. http://www.arpa.mil/ipto/programs/gale/index.htm.
[3] :- Technology and Corpora for Speech to Speech Translation. http://www.tc-star.org/.
[4] EACL 09 Fourth Workshop on Statistical Machine Translation. http://www.statmt.org/wmt09/.

REFERENCES

A. Ahmed and G. Hanneman. 2005. *Syntax-based Statistical Machine Translation: A Review*. Technical Report. Carnegie Mellon University. Retrieved from http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/cmt-55/lti/Courses/734/Spring-08/Amr%2BGreg-survey-SSMT.pdf

Y. Al-Onaizan and K. Knight. 2002. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 400–408. DOI:http://dx.doi.org/10.3115/1073083.1073150

H. Alshawi, S. Douglas, and S. Bangalore. 2000. Learning dependency translation models as collections of finite-state head transducers. *Comput. Linguist.* 26, 1 (March 2000), 45–60. DOI:http://dx.doi.org/10.1162/089120100561629

A. Aue, A. Menezes, B. Moore, C. Quirk, and E. Ringger. 2004. Statistical Machine Translation Using Labeled Semantic Dependency Graphs. In *Proceedings of TMI 2004*. 125–134.

E. Avramidis and P. Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics and Human Language Technology (ACL-HLT'08)*. Association for Computational Linguistics, Stroudsburg, PA, 763–770.

A. Aw, M. Zhang, J. Xiao, and J. Su. 2006. A phrase-based statistical model for SMS text normalization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computer Linguistics, Stroudsburg, PA. DOI:http://dx.doi.org/P/P06/P06-2005.pdf

N. Bach. 2012. *Dependency Structures for Statistical Machine Translation*. PhD dissertation. Carnegie Mellon University.

I. Badr, R. Zbib, and J. Glass. 2009. Syntactic phrase reordering for English-to-Arabic statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*. Association for Computational Linguistics, Stroudsburg, PA, 86–93.

L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the Linguistic Annotation Workshop*. Association for Computational Linguistics, Stroudsburg, PA.

R. E. Banchs and M. R. Costa-jussà. 2011. A semantic feature for statistical machine translation. In *Proceedings of the 5th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5)*. Association for Computational Linguistics, Stroudsburg, PA, 126–134.

S. Bangalore, P. Haffner, and S. Kanthak. 2007. Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*. Association for Computational Linguistics, Stroudsburg, PA, 152–159.

A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22, 1 (March 1996), 39–72.

N. Bertoldi, M. Cettolo, and M. Federico. 2010. Statistical machine translation of texts with misspelled words. In *Proceedings of the NAACL*. 412–419.

J. A. Bilmes and K. Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *Proceedings of the Conference of the Association for Computational Linguistics and Human Language Technology (NAACL-HLT'03)*. Association for Computational Linguistics, Stroudsburg, PA, 4–6.

A. Birch and M. Osborne. 2010. LRscore for evaluating lexical and reordering quality in MT. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR (WMT'10)*. Association for Computational Linguistics, Stroudsburg, PA, 327–332.

A. Birch, M. Osborne, and P. Koehn. 2007. CCG Supertags in Factored Translation Models. In *Proceedings of the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA.

H. C. Boas. 2002. Bilingual FrameNet dictionaries for machine translation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. 1364–1371.

O. Bojar, M. Ercegovčević, M. Popel, and O. Zaidan. 2011. A grain of salt for the WMT manual evaluation output. In *Proceedings of the EMNLP 6th Workshop on Statistical Machine Translation (WMT'11)*. 1–11.

O. Bojar and A. Tamchyna. 2011. Forms wanted: Training SMT on monolingual Data. In *Proceedings of the Workshop of Machine Translation and Morphologically-Rich Languages*.

T. Brants. 2000. A statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference*.

P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19, 2 (1993), 263–311.

M. Carpuat and D. Wu. 2007. Context-dependent phrasal translation lexicons for statistical machine translation. In *Proceedings of the Machine Translation Summit XI*.

M. Carpuat and D. Wu. 2008. Evaluation of Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*.

Y. S. Chan, H. T. Ng, and D. Chiang. 2007. Word Sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL'07)*. Association for Computational Linguistics, Stroudsburg, PA, 33–40.

P. Charoenpornsawat, V. Sornlertlamvanich, and T. Charoenporn. 2002. Improving translation quality of rule-based machine translation. In *Proceedings of the 2002 COLING Workshop on Machine translation in Asia, Volume 16 (COLING-MTIA'02)*. Association for Computational Linguistics, Stroudsburg, PA, 1–6. DOI:http://dx.doi.org/10.3115/1118794.1118799

S. F. Chen and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics (ACL'96)*. Association for Computational Linguistics, Stroudsburg, PA, 310–318. DOI:http://dx.doi.org/10.3115/981863.981904

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, Stroudsburg, PA, 263–270.

D. Chiang. 2007. Hierarchical phrase-based translation. *Comput. Linguist.* 33, 2 (June 2007), 201–228. DOI:http://dx.doi.org/10.1162/coli.2007.33.2.201

D. Chiang, K. Knight, and W. Wang. 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 (NAACL'09)*. Association for Computational Linguistics, Stroudsburg, PA, 218–226.

M. Collins, P. Koehn, and I. Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the Annual Conference of the Association for Computational Lingusitics (ACL'05)*. Association for Computational Linguistics, Stroudsburg, PA.

M. R. Costa-Jussà. 2012. An overview of the phrase-based statistical machine translation techniques. *Knowledge Eng. Review* 27, 4 (2012), 413–431.

M. R. Costa-jussà, R. E. Banchs, E. Rapp, P. Lambert, K. Eberle, and B. Babych. 2013. Workshop on hybrid approaches to translation: Overview and developments. In *Proceedings of the ACL 2nd Workshop on Hybrid Approaches to Translation (HyTra'13)*. Association for Computational Linguistics, Stroudsburg, PA.

M. R. Costa-Jussà and J. A. R. Fonollosa. 2009. State-of-the-art word reordering approaches in statistical machine translation: A survey. *IEICE Transactions on Information and Systems* 92, 11 (2009), 2179–2185.

B. A. Cowan. 2008. *A Tree-to-Tree Model for Statistical Machine Translation*. Ph.D. Dissertation. Standford University.

M. Creutz and K. Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*.

A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne. 2009. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, Stroudsburg, PA, 73–76.

M. Diab, M. Ghoneim, and N. Habash. 2007. Arabic diacritization in the context of statistical machine translation. In *Proceedings of the Machine Translation Summit XI*. 143–149.

Y. Ding and M. Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, Stroudsburg, PA, 541–548. DOI:http://dx.doi.org/10.3115/1219840.1219907

A. Eisele, C. Federmann, H. Saint-Amand, M. Jellinghaus, T. Herrmann, and Y. Chen. 2008. Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, 179–182.

I. D. El-Kahlout and K. Oflazer. 2010. Exploiting morphology and local wword reordering in English-to-Turkish phrase-based statistical machine translation. *IEEE Transactions on Audio, Speech & Language Processing* 18, 6 (2010), 1313–1322.

A. El Kholy and N. Habash. 2012. Orthographic and morphological processing for English-Arabic statistical machine translation. *Machine Translation* 26, 1–2 (2012), 25–45. DOI:http://dx.doi.org/10.1007/s10590-011-9110-0

J. Elming. 2008. *Syntactic Reordering in Statistical Machine Translation*. PhD dissertation. Copenhaguen Business School.

C. España-Bonet, J. Giménez, and L. Màrquez. 2009. Discriminative phrase-based models for Arabic machine yranslation. *ACM Transactions on Asian Language Information Processing Journal* 8, 4 (March 2009), Article 15. 20 pages. DOI:http://dx.doi.org/10.1145/1644879.1644882

C. España-Bonet, G. Labaka, A. D. de Ilarraza, L. Màrquez, and K. Sarasola. 2011. Hybrid Machine Translation Guided by a Rule-Based System. In *Proceedings of the 13th Machine Translation Summit.* 554–561.

M. Farrús, M. R. Costa-Jussà, J. B. Marino, M. Poch, A. Hernandez, C. Henríquez, and J. A. R. Fonollosa. 2011. Overcoming statistical machine translation limitations: Error analysis and proposed solutions for the Catalan-Spanish language pair. *Language Resources and Evaluation* (2011), 181–208.

M. Farrús, M. R. Costa-jussà, J. B. Marino, and J. A. R. Fonollosa. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT'10)*. 167–173.

M. Farrús, M. R. Costa-jussà, and M. Popović. 2012. Study and correlation analysis of linguistic, perceptual, and automatic machine translation evaluations. *J. Am. Soc. Inf. Sci. Technol.* 63, 1 (Jan. 2012), 174–184. DOI:http://dx.doi.org/10.1002/asi.21674

M. Felice and L. Specia. 2012. Linguistic features for quality estimation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, 96–103.

M. Flanagan. 1994. Error classification for MT evaluation. In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas* (1994), 65–72.

M. L. Forcada, F. M. Tyers, and G. Ramírez-Sánchez. 2009. The Apertium machine translation platform: Five years on. In *Proceedings of the 1st International Workshop on Free/Open-Source Rule-Based Machine Translation*, Juan Antonio Prez-Ortiz, Felipe Snchez-Martnez, and Francis M. Tyers (Eds.). Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, 3–10.

Ll. Formiga, A. Hernández, J. B. Mariño, and E. Monte. 2012. Improving English to Spanish out-of-domain translations by morphology generalization and generation. In *Proceedings of the AMTA Workshop on Monolingual Machine Translation*.

G. Foster, P. Isabelle, and R. Kuhn. 2010. Translating structured documents. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas*.

A. Fraser and D. Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics* (2007), 293–303.

P. Fung and P. Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and E. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*. 57–63.

M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL-44)*. Association for Computational Linguistics, Stroudsburg, PA, 961–968. DOI:http://dx.doi.org/10.3115/1220175.1220296

M. Galley, M. Hopkins, K. Knight, and D. Marcu. 2004. What's in a translation rule? In *Proceedings of the 2004 Annual Conference of the North American Chapter of the Association for Computational Linsuitics (NAACL HLT 2004)*, Daniel Marcu Susan Dumais and Salim Roukos (Eds.). Association for Computational Linguistics, Stroudsburg, PA, 273–280.

I. García-Varea, F. J. Och, H. Ney, and F. Casacuberta. 2001. Refined lexicon models for statistical machine translation using a maximum entropy approach. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter of the ASsociation for Computational Linguistics (ACL/EACL'01)*. Association for Computational Linguistics, Stroudsburg, PA.

D. Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*. Association for Computational Linguistics, Stroudsburg, PA, 376–384.

U. Germann. 2012. Syntax-aware phrase-based statistical machine translation: System description. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, 292–297.

J. Giménez and L. Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (StatMT'07)*. Association for Computational Linguistics, Stroudsburg, PA, 256–264.

J. Graehl, K. Knight, and J. May. 2008. Training tree transducers. *Comput. Linguist.* 34, 3 (Sept. 2008), 391–427. DOI:http://dx.doi.org/10.1162/coli.2008.07-051-R2-03-57

S. Green and J. DeNero. 2012. A Class-based agreement model for generating accurately inflected translations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA.

R. Haque. 2011. *Integrating Source-Language Context into Log-linear Models of Statistical Machine Translation*. Ph.D. Dissertation. Dublin City University.

C. Hardmeier. 2012. Discourse in Statistical Machine Translation: A Survey and a Case Study. *Discours* 11 (2012). http://discours.revues.org/8726.

C. Hardmeier and M. Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT'10)*, Marcello Federico, Ian Lane, Michael Paul, and François Yvon (Eds.). 283–289.

R. R. Hausser. 2001. *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*. Springer.

S. Helmreich and D. Farwell. 1998. Translation differences and pragmatics-based MT. *Machine Translation* 13, 1 (1998), 17–39. DOI:http://dx.doi.org/10.1023/A:1008062303478

H. Hoang and A. Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT'09)*. 152–159.

C. Huang, H. Yen, P. Yang, S. Huang, and J. S. Chang. 2011. Using sublexical translations to handle the OOV problem in machine translation. 10, 3, Article 16 (Sept. 2011), 20 pages. DOI:http://dx.doi.org/10.1145/2002980.2002986

L. Huang, K. Knight, and A. Joshi. 2006. A syntax-directed translator with extended domain of locality. In *Proceedings of the Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing (CHSLP'06)*. Association for Computational Linguistics, Stroudsburg, PA, 1–8.

W. J. Hutchins. 1995. Machine translation: A brief history. In *Concise History of the Language Sciences: From the Sumerians to the Cognitivists.* Pergamon Press, 431–445.

W. J. Hutchins. 2005. The History of Machine Translation in a Nutshell. Retrieved from http://ourworld.compuserve.com/homepages/WJHutchins/Nutshell.htm.

V. Istvan and Y. Shoichi. 2009. Bilingual dictionary generation for low-resourced language pairs. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Vol. 2. 862–870.

P. Karageorgakis, A. Potamianos, and K. Ioannis. 2005. Towards incorporating language morphology into statistical machine translation systems. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop.*

M. Khalilov and J. A. R. Fonollosa. 2011. Syntax-based reordering for statistical machine translation. *Computer Speech and Language Journal* 25, 4 (October 2011).

R. Kneser and H. Ney. 1995. Improved backing-off for n-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 49–52.

K. Knight and J. Graehl. 1998. Machine transliteration. *Comput. Linguist.* 24, 4 (Dec. 1998), 599–612.

Catherine Kobus, François Yvon, and Géraldine Damnati. 2008. Normalizing SMS: Are two metaphors better than one? In *Proceedings of the 22nd International Conference on Computational Linguistics, Proceedings of the Conference (COLING'08).* 441–448. DOI:http://dx.doi.org/anthology/C08-1056

P. Koehn and H. Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*. Association for Computational Linguistics, Stroudsburg, PA, 868–876.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL'07)*. Association for Computational Linguistics, Stroudsburg, PA, 177–180.

P. Koehn and K. Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics, Volume 1*

(EACL'03). Association for Computational Linguistics, Stroudsburg, PA, 187–193. DOI:http://dx.doi.org/10.3115/1067807.1067833

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Annual Conference of the Association for Computational Lingusitics (ACL03)*. Association for Computational Linguistics, Stroudsburg, PA, USA.

G. Kondrak. 2005. Cognates and word alignment in bitexts. In *Proceedings of the 10th Machine Translation Summit*. 305–312.

G. Kondrak, D. Marcu, and K. Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–Short Papers, Volume 2 (NAACL-Short'03)*. Association for Computational Linguistics, Stroudsburg, PA, 46–48. DOI:http://dx.doi.org/10.3115/1073483.1073499

A. Kumaran and T. Kellner. 2007. A generic framework for machine transliteration. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*. ACM, New York, NY, 721–722. DOI:http://dx.doi.org/10.1145/1277741.1277876

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML'01)*. Morgan Kaufmann, San Francisco, CA, 282–289.

T. K. Landauer, D. Laham, and P. Foltz. 1998. Learning human-like knowledge by singular value decomposition: A progress report. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 45–51.

P. Langlais and F. Gotti. 2006. Phrase-based SMT with shallow tree-phrases. In *Proceedings of the Workshop on Statistical Machine Translation (StatMT'06)*. Association for Computational Linguistics, Stroudsburg, PA, 39–46.

P. Langlais and A. Patry. 2007. Translating unknown words by analogical learning. In *EMNLP-CoNLL* (2010-06-04). ACL, 877–886.

A. Lavie and A. Agarwal. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (StatMT'07)*. Association for Computational Linguistics, Stroudsburg, PA, 228–231.

R. L. Nagard and P. Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR'10)*. Association for Computational Linguistics, Stroudsburg, PA, 258–267.

C. Li, N. Duan, Y. Zhao, S. Liu, L. Cui, M. Hwang, A. Axelrod, J. Gao, Y. Zhang, and L. Deng. 2010. The MSRA machine translation system for IWSLT 2010. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT'10)*, 135–138.

C. Li, D. Zhang, M. Li, M. Zhou, M. Li, and Y. Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of the Annual Conference of the Association for Computational Lingusitics (ACL'07)*. Association for Computational Linguistics, Stroudsburg, PA, 720–727.

Z. Li and D. Yarowsky. 2008. Unsupervised translation induction for Chinese abbreviations using monolingual corpora. In *ACL*, Kathleen McKeown, Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui (Eds.). Association for Computer Linguistics, Stroudsburg, PA, 425–433.

L. V. Lita, A. Ittycheriah, S. Roukos, and N. Kambhatla. 2003. tRuEcasIng. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. 152–159.

Y. Liu, Q. Liu, and S. Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL-44)*. Association for Computational Linguistics, Stroudsburg, PA, 609–616. DOI:http://dx.doi.org/10.3115/1220175.1220252

C. Lo and D. Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (HLT'11)*. Association for Computational Linguistics, Stroudsburg, PA, 220–229.

LSA. 2013. Linguistic Society of America Homepage. Retrieved from http://www.linguisticsociety.org.

J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, and M. R. Costa-jussà. 2006. N-gram-based Machine Translation. *Comput. Linguist.* 32, 4 (Dec. 2006), 527–549. DOI:http://dx.doi.org/10.1162/coli.2006.32.4.527

Y. Marton, C. Callison-Burch, and P. Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1 (EMNLP'09)*. Association for Computational Linguistics, Stroudsburg, PA, 381–390.

I. A. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard. 2004. *On the Use of Information Retrieval Measures for Speech Recognition Evaluation*. Idiap-RR Idiap-RR-73-2004. IDIAP, Martigny, Switzerland.

A. Menezes and C. Quirk. 2008. Syntactic models for structural word insertion and deletion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*. Association for Computational Linguistics, Stroudsburg, PA, 735–744.

A. Menezes and S. D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Workshop on Data-driven Methods in Machine Translation, Volume 14 (DMMT'01)*. Association for Computational Linguistics, Stroudsburg, PA, 1–8. DOI:http://dx.doi.org/10.3115/1118037.1118043

T. Meyer, A. Popescu-Belis, N. Hajlaoui, and A. Gesmundo. 2012. Machine translation of labeled discourse connectives. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas (AMTA'12)*. Retrieved from http://www.mt-archive.info/AMTA-2012-Meyer.pdf.

E. Minkov, K. Toutanova, and H. Suzuki. 2007. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA.

S. Mirkin, L. Specia, N. Cancedda, I. Dagan, M. Dymetman, and I. Szpektor. 2009. Source-language entailment modeling for translating unknown terms. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 (ACL'09)*. Association for Computational Linguistics, Stroudsburg, PA, 791–799. http://dl.acm.org/citation.cfm?id=1690219.1690257

R. Mitkov, V. Pekar, D. Blagoev, and A. Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation* 21, 1 (March 2007), 29–53. DOI:http://dx.doi.org/10.1007/s10590-008-9034-5

A. Mulloni and A. Pekar. 2006. Automatic detection of orthographic cues for cognate recognition. In *Proceedings of the Conference on Language Resources and Evaluation*.

P. Nakov and H. T. Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*. ACL, 1358–1367. DOI:http://dx.doi.org/anthology/D09-1141

F. J. Och. 2003. Minimum Error Rate Training In Statistical Machine Translation. In *41th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, 160–167.

F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, 295–302.

F. J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Comput. Linguist.* 30, 4 (Dec. 2004), 417–449. DOI:http://dx.doi.org/10.1162/0891201042544884

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*. Association for Computational Linguistics, Stroudsburg, PA, 311–318. DOI:http://dx.doi.org/10.3115/1073083.1073135

M. Popović, A. de Gispert, D. Gupta, P. Lambert, H. Ney, J. B. Mariño, M. Federico, and R. Banchs. 2006. Morpho-syntactic information for automatic rrror analysis of statistical machine translation output. In *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, 1–6.

M. Popović and H. Ney. 2007. Word error rates: Decomposition over POS classes and applications for error analysis. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (StatMT'07)*. Association for Computational Linguistics, Stroudsburg, PA, 48–55.

M. Popović and H. Ney. 2009. Syntax-oriented evaluation measures for machine translation output. In *Proceedings of the 4th Workshop on Statistical Machine Translation (StatMT'09)*. Association for Computational Linguistics, Stroudsburg, PA, 29–32.

M. Popović and H. Ney. 2011. Towards automatic error analysis of machine translation output. *Comput. Linguist.* 37, 4 (Dec. 2011), 657–688. DOI:http://dx.doi.org/10.1162/COLI_a_00072

C. Quirk, A. Menezes, and C. Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, Stroudsburg, PA, 271–279. DOI:http://dx.doi.org/10.3115/1219840.1219874

A. Razmara. 2011. *Application of Tree Transducers in Statistical Machine Translation*. Technical Report. Depth Report, Simon Fraser University.

J. Riesa, B. Mohit, K. Knight, and D. Marcu. 2006. Building an English-Iraqi Arabic machine translation system for spoken utterances with limited resources. In *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH'06)*.

E. Ringger, M. Gamon, R. C. Moore, D. Rojas, M. Smets, and S. Corston-Oliver. 2004. Linguistically informed statistical models of constituent structure for ordering in sentence realization. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*. Association for Computational Linguistics, Stroudsburg, PA, article 673. DOI:http://dx.doi.org/10.3115/1220355.1220452

R. Rosa, D. Mareček, and O. Dušek. 2012. DEPFIX: A system for automatic correction of Czech MT outputs. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, 362–368.

G. Salton and M. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

L. Shao and H. T. Ng. 2004. Mining new word translations from comparable corpora. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*. Association for Computational Linguistics, Stroudsburg, PA, article 618. DOI:http://dx.doi.org/10.3115/1220355.1220444

L. Shen, J. Xu, and R. Weischedel. 2010. String-to-dependency statistical machine translation. *Comput. Linguist.* 36, 4 (Dec. 2010), 649–671. DOI:http://dx.doi.org/10.1162/coli_a_00015

L. Shen, B. Zhang, S. Matsoukas, and R. Weischedel. 2009. Effective use of linguistic and contextual information for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*. Association for Computational Linguistics, Stroudsburg, PA, 72–80.

M. Simard, N. Cancedda, B. Cavestro, M. Dymetman, E. Gaussier, C. Goutte, K. Yamada, P. Langlais, and A. Mauser. 2005. Translating with non-contiguous phrases. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05)*. Association for Computational Linguistics, Stroudsburg, PA, 755–762. DOI:http://dx.doi.org/10.3115/1220575.1220670

D. A. Smith and J. Eisner. 2006. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings of the Workshop on Statistical Machine Translation (StatMT'06)*. Association for Computational Linguistics, Stroudsburg, PA, 23–30.

M. G. Snover, N. Madnani, B. Dorr, and R. Schwartz. 2009. TER-Plus: Paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation* 23, 2–3 (Sept. 2009), 117–127. DOI:http://dx.doi.org/10.1007/s10590-009-9062-9

S. Stymne. 2011. BLAST: A tool for error analysis of machine translation output. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*. Association for Computer Linguistics, 56–61.

2009 Thrumair. 2009. Comparing different architectures of hybrid machine translation systems. In *Proceedings of the MT-Summit XII*.

C. Tillman. 2004. A block orientation model for statistical machine translation. In *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL'04)*. Association for Computational Linguistics, Stroudsburg, PA.

J. P. Turian, B. Wellington, and I. D. Melamed. 2006. Scalable Discriminative learning for natural language parsing and translation. In *Proceedings of the 2006 Neural Information Processing Systems (NIPS'06)*. Bernhard Schlkopf, John Platt, and Thomas Hoffman (Eds.). MIT Press, 1409–1416.

N. Ueffing and H. Ney. 2003. Using POS information for statistical machine translation into morphologically rich languages. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics (EACL'03)*. Association for Computational Linguistics, Stroudsburg, PA, 347–354.

A. Venugopal and A. Zollmann. 2009. Grammar based statistical MT on Hadoop: An end-to-end toolkit for large scale PSCFG based MT. In *The Prague Bulletin of Mathematical Linguistics No. 91*. 67–78.

D. Vilar. 2011. *Investigations on Hierarchical Phrase-based Machine Translation*. Ph.D. Dissertation. RWTH Aachen University, Aachen, Germany.

D. Vilar, J. Xu, L. Fernando-D'Haro, and H. Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'06)*. 697–702.

P. Virga and S. Khudanpur. 2003. Transliteration of proper names in cross-lingual information retrieval. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, Volume 15 (MultiNER'03)*. Association for Computational Linguistics, Stroudsburg, PA, 57–64. DOI:http://dx.doi.org/10.3115/1119384.1119392

S. Virpioja, J. J. Väyrynen, M. Creutz, and M. Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of the Machine Translation Summit XI*. 491–498.

C. Wang, M. Collins, and P. Koehn. 2007. Chinese syntactic reordering for statistical machinetranslation. In *Empirical Methods in Natural Language Processing (EMNLP'07)*. Association for Computational Linguistics, Stroudsburg, PA.

W. Wang, K. Knight, and D. Marcu. 2006. Capitalizing machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. Association for Computational Linguistics, New York, NY, 1–8.

B. Webber. 2012. Discourse and SMT: Where and How? (Sept. 2012). Seventh Machine Translation Marathon 2012. Invited talk.

D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.* 23, 3 (Sept. 1997), 377–403.

D. Wu. 2009. Toward machine translation with statistics and syntax and semantics. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU'09)*. 12–21.

D. Wu and P. Fung. 2009. Semantic roles for SMT: A hybrid two pass model. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT'09)*. Association for Computational Linguistics, Stroudsburg, PA.

F. Xia and M. McCord. 2004. Improving a Statistical MT System with Automatically Learned Rewrite Patterns. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*. Association for Computational Linguistics, Stroudsburg, PA.

K. Yamada and K. Knight. 2002. A decoder for syntax-based statistical MT. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*. Association for Computational Linguistics, Stroudsburg, PA, 303–310. DOI:http://dx.doi.org/10.3115/1073083.1073134

R. Zens, F. J. Och, and H. Ney. 2002. Phrase-Based Statistical Machine Translation. In *Proceedings of the German Conference on Artificial Intelligence (KI'02)*. Springer-Verlag.

H. Zhang and D. Gildea. 2005. Stochastic lexicalized inversion transduction grammar for alignment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 475–482. DOI:http://dx.doi.org/10.3115/1219840.1219899

J. Zhang, F. Zhai, and C. Zhing. 2012. Handling unknown words in statistical machine translation from a new perspective. In *Proceedings of the NLPCC*.

M. Zhang, A. Aw H. Jiang, J. Sun, S. Li, and C. Tan. 2007. A tree-to-tree alignment-based model for SMT. In *Proceedings of the MT-Summit*. 535–542.

M. Zhang, H. Li, and J. Su. 2004. Direct orthographical mapping for machine transliteration. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*. Association for Computational Linguistics, Stroudsburg, PA, article 716. DOI:http://dx.doi.org/10.3115/1220355.1220458

Received March 2013; revised June 2013; accepted September 2013