 ISBN: 978-84-608-1989-9

Hierarchical Phrase-based Translation Model vs. Classical Phrase-based Translation Model for Spanish-English Statistical Machine Translation System

Modelo de Traducción Jerárquico basado en Frases vs. Modelo de Traducción Clásico basado en Frases en un Sistema de Traducción Automática Estadística para Español-Inglés

Benyamin Ahmadnia
Autonomous University of Barcelona
08193 Cerdanyola del Valles, Spain
benyamin.ahmadnia@uab.cat

Javier Serrano
Autonomous University of Barcelona
08193 Cerdanyola del Valles, Spain
javier.serrano@uab.cat

Resumen: La traducción automática estadística es un método que adquiere conocimiento automáticamente a partir de grandes cantidades de datos de entrenamiento. Existen varias aproximaciones para el entrenamiento de sistemas de traducción automática estadística, tales como basada en palabras, basada en frases, basada en sintaxis y jerárquica basada en frases. En este trabajo comparamos un modelo de traducción clásico basado en frases y un modelo de traducción jerárquico basado en frases en la dirección de la traducción Español-Inglés, y su dirección contraria. Se demuestra que un sistema de traducción jerárquico basado en frases superará a un sistema clásico basado en frases en la dirección de traducción Español-Inglés, pero para la dirección Inglés-Español, el sistema basado en frases clásico es preferible. Buscamos explicar los detalles de los experimentos de traducción con nuestro sistema de traducción automática estadística a partir de datos paralelos, con las herramientas Moses (no jerárquica) y cdec (jerárquica).
Palabras clave: Traducción automática estadística, procesamiento del lenguaje natural, el modelo basado en la frase clásica, modelo basado en frases jerárquica

Abstract: Statistical machine translation is a method that automatically acquires knowledge from large amounts of training data. There are some approaches in order to train a statistical machine translation system such as word-based, phrase-based, syntax-based, and hierarchical phrase-based. In this paper we compare a classical phrase-based translation model and a hierarchical phrase-based translation model in Spanish-English translation direction, and back translation. We show that a hierarchical phrase-based translation system will outperform a classical phrase-based system in the Spanish-English translation direction, but for the English-Spanish direction, the classical phrase-based system is preferable. We seek to explain the detail of translation experiments with our statistical machine translation system using parallel data, with Moses (non-hierarchical) and cdec (hierarchical) toolkits.
Keywords: Statistical machine translation, natural language processing, classical phrase-based model, hierarchical phrase-based model

1 Introduction

Natural Language Processing (NLP) is a barnch of artificial intelligence and theory-motivated range of computational techniques for the automatic analysis and representation of human language.The representation of human language is defined on certain levels of linguistic analysis for achieving the human-like processing. The goal of NLP is to accomplish unambiguous human-like language processing (White and Cambria, 2014).

Machine Translation (MT) is one of the earliest areas of research in NLP. It is the automatic translation from one natural language into another using computers. MT refers to computerized systems responsible for the production of translations with or without human assistance. It excludes computer-based translation tools which support translators by providing access to on-

line dictionaries, remote terminology databanks, transmission and reception of texts, etc (Hutchins, 1995).

Research works in this field dates as far back as the 1950's. Several different translation methods have been explored to date, the oldest and perhaps the simplest method being rule-based translation, which is in reality transliteration, or translating each word in the source language with its equivalent counterpart in the target language. This method is very limited in the accuracy it can give. A method known as Statistical Machine Translation (SMT) seems to be the preferred approach of many industrial and academic research laboratories, due to its recent success. Different evaluation metrics generally show statistical machine translation approaches to yield higher scores (Koehn, Och, and Marcu, 2003).

SMT requires enormous amounts of parallel text in the source and target language to achieve high quality translation. However, many languages are considered to be low-density languages, because the population speaking the language is not very large.The goal of SMT is to translate a source language sequence into a target language sequence by maximizing the posterior probability of the target sequence given the source sequence. In state-of-the-art translation systems, this posterior probability usually is modelled as a combination of several different models, such as phrase-based models for both translation directions, lexicon models for both translation directions, target language model, phrase and word penalties, etc. Translation model probabilities that describe correspondences between the words in the source language and the words in the target language are learned from a bilingual parallel text corpus and language model probabilities are learned from a monolingual text in the target language (Marcu and Wong, 2002).

Most recent research in the area of statistical machine translation has been targeted at modelling translation based on phrases in both the source, and the target languages. Many modern successful translation machines use this translation approach (Och and Ney, 2004).

In this paper we compare the classical phrase-based translation model with the hierarchical phrase-based translation model in Spanish-English translation direction, and back translation, using the BLEU and the TER as the evaluation metrics in order to compare the results between classical and hierarchical models.

1.1 Classical vs. Hierarchical SMT

One of the approaches of SMT is word-based translation. As the name suggests, the words in an input sentence are translated word by word individually, and these words finally are arranged in a specific way to get the target sentence. This approach is the very first attempt in SMT systems technology that is comparatively simple and efficient. The main disadvantage of this system is the oversimplified word by word translation of sentences, which may reduce the performance of the translation system.

In order to reduce the limitation of this approach, phrase-based translation approach introduced, where each source and target sentence is divided into separate phrases instead of words before translation. The alignment between the phrases in the input and output sentences normally follows certain patterns, which is very similar to word-based translation. Even though the phrase-based models result in better performance than the word-based translation, they did not improve the model of sentence order patterns. The alignment model is based on classical reordering patterns, and experiments show that this reordering technique may perform well with local phrase orders but not as well with long sentences and complex orders.

By considering the drawback of previous two methods developed a more sophisticated SMT approach, another model introduced which called hierarchical phrase-based translation model (Chiang, 2005). The advantage of this approach is that, hierarchical phrases have recursive structures instead of simple phrases. This higher level of abstraction approach further improved the accuracy of the SMT system.

Machine translation can be divided into three steps: training the translation model, tuning parameters, and decoding. We will mostly focus on the first step, since that is where classical and hierarchical MT approaches differ the most.

The output of the first step is the translation model. For both classical and hierarchical variants, the translation model consists of

a set of rules in the following format:

$$\alpha = \alpha_0\alpha_1...||\beta = \beta_0\beta_1...||\chi||\ell(\alpha \rightarrow \beta) \quad (1)$$

We call the sequence of α_i's the source side of the rule, and sequence of β_j's the target side of the rule. The above indicates that the source side translates into the target side with a likelihood of $\ell(\alpha \rightarrow \beta)$[1]. χ contains token alignments in the format $i-j$, indicating that source token α_i is aligned to target token β_j.

A hierarchical model differs from a classical model in terms of rule expressivity: rules are allowed to contain one or more non-terminals, each acting as a variable that can be expanded into other expressions using the grammar, carried out in a recursive fashion. These grammars are called synchronous context-free grammars (SCFG), as each rule describes a context-free expansion on both sides. Consider the following two rules from an SCFG:
I)[X] leave in europe || permiso [X] en europa || 1-0 2-3 3-4 || 1
II)maternity || maternidad || 0-0 || 0.69
In I, the non-terminal variable [X] allows an arbitrarily long part of the sentence to be moved from the left of the sentence in English to the middle of the sentence in Spanish, even though it generates a single token using II in this particular example. As a result, an SCFG can capture distant dependencies in language that may not be realized in classical models.

Each sequence of rules that covers the entire input is called a derivation, D, and produces a translation candidate, t, which is scored by a linear combination of features. One can use many features to score a candidate, but two features are the most important: the product of rule likelihood values indicates how well the candidate preserves the original meaning, $TM(t, D|s)$, whereas the language model score, $LM(t)$, indicates how well-formed the translation is. Combining the two, the decoder searches for the best translation:

$$t = \arg\max_t \{max\ TM(t, D|s)LM(t)\} \quad (2)$$

There is a tradeoff between using either classical or hierarchical grammars. The latter provides more expressivity in representing linguistic phenomena, but at the cost of slower decoding. On the other hand, classical models are faster, but less expressive. Also, due to the lack of variables, classical grammars contain more rules, resulting in a more verbose translation grammar (Ture and Lin, 2013).

A significantly critical task in a classical phrase-based SMT system is the determination of a translation model from a word-aligned parallel corpus. A phrase table containing the source language phrases, their target language equivalents and their associated probabilities, in most systems is extracted in a preprocessing stage before decoding a test set (Koehn, Och, and Marcu, 2003).

In the hierarchical phrase-based approach, translation is modelled by using SCFGs. In general, probabilistic SCFGs can be learned from word-aligned parallel data using heuristic methods (Chiang, 2007). We can first extract initial phrase pairs and then obtain hierarchical phrase rules. Once the SCFG is obtained, new sentences can be decoded by finding the most likely derivation of SCFG rules. The Hiero SCFG allows vast numbers of derivations which can make unconstrained decoding intractable.

1.2 Toolkits

There are some open source engines for training machine translation such as Joshua, Apertium, Moses and cdec.

Joshua is a general-purpose open source toolkit used for parsing-based machine translation, accomplishing the same purpose as Moses toolkit does for regular phrase-based machine translation. It is written in Java programming language.

Apertium is an open source machine translation system for the languages of Spain which is funded by the Spanish government. It is designed to translate between closely related languages, although it has recently been expanded to treat more divergent language pairs. To create a new machine translation system, one just has to develop linguistic data in well-specified XML formats.

We use two different open-source toolkits for our statistical machine translation system whose contributions are: support for linguistically motivated factors, confusion network decoding, and efficient data formats for trans-

[1]The likelihood function is not a probability density function because it is not normalized.

lation models and language models. In addition to statistical machine translation decoder, the toolkits also include a wide variety of tools for training, tuning and applying the system to many translation tasks.

We use cdec because it is a mature software platform for research in development of translation models and algorithms. Its architecture was developed with machine learning and algorithmic research use-cases in mind. It is designed to run efficiently in both limited resource environments (single processor, limited memory) up to very large cluster environments. Also we use Moses because that allows us to automatically train translation models for our considered language pair. Once we have a trained model, an efficient search algorithm quickly finds the highest probability translation among the exponential number of choices.

We conducted experiments with hierarchical translation models using cdec, with a range of corpora sizes, and compared the results with classical phrase-based models using Moses with the same corpora.

1.2.1 Moses Toolkit

Moses (Koehn et al., 2007), is an open source phrase-based toolkit that can be used to train statistical models of text translation from a source language to a target one. Moses allows new source-language text to be decoded using these models to produce automatic translations in the target language. Training requires a parallel corpus of passages in the two languages, typically manually translated sentence pairs.

A subsample of occurrences of given source phrase are used to calculate translation probabilities. Phrase translations and their model parameters can be determined at runtime as the system accesses the target language corpus and word alignment data. A suffix array can also be used to obtain hierarchical phrases at run-time (Lopez, 2008).

1.2.2 cdec Toolkit

cdec (Dyer et al., 2010), is a decoder, aligner, and learning framework for statistical machine translation and similar structured prediction models.

It is using a single unified internal representation for translation forests, the decoder strictly separates model-specific translation logic from general re-scoring, pruning, and inference algorithms. From this unified representation, the decoder can extract not only the 1-best or k-best translations, but also alignments to a reference, or the quantities necessary to drive discriminative training using gradient-based or gradient-free optimization techniques. Its efficient C++ implementation means that memory use and run-time performance are significantly better than comparable decoders.

1.3 Evaluation Metrics

One aspect of machine translation that poses a challenge is developing an effective automated metric for evaluating machine translation. This is because each output sentence has a number of acceptable translations.

The main idea of automated evaluation is comparing output of a machine translation system to a good reference (usually human) translation. It can be used on an ongoing basis during system development to test changes. Also it is fast and cheap, with minimal human labor, and it is not necessary to use bilingual speakers.

Most popular metrics yield scores primarily based on matching phrases in the translation produced by the system to those in several reference translations. The metric scores mostly differ in how they show reordering and synonyms. The metrics we chose to work with, are BLEU and TER, while BLEU is a relatively simple metric, it has a number of shortcomings. For BLUE interpretation scores, the higher score, and for TER interpretation scores, the lowest score are suitable in order to compare the translation systems.

1.3.1 BLEU

Bilingual Evaluation Understudy (BLEU) is an MT evaluation technique that is quick, inexpensive, and language independent.

BLEU is one of the metrics to achieve a high correlation with human judgements of quality and remains one of the most popular automated and inexpensive metrics. In general, BLEU is the most popular metrics used for both comparison of translation systems and tuning of machine translation models (Papineni et al., 2001).

BLEU uses a modified form of N-gram precision to compare a candidate translation with multiple reference translations. It applies a length penalty (brevity penalty) if the generated sentence is shorter than the best matching (in length) reference translation.

1.3.2 TER

Translation Error Rate (TER) is an extension of Word Error Rate (WER) [2]. It is an other error metric for MT that operates by measuring the amount of editing that a human would have to undertake to produce a translation so that it forms an exact match with a reference translation (Snover et al., 2006).

This technique is a more intuitive measure of goodness of machine translation output-specifically, the number of edits needed to fix the output so that, it semantically matches a correct translation. Human targeted TER yields higher correlations with human judgement than BLEU.

2 Data Preparation

In order to provide the best possible results, a statistical language model requires an extremely large amounts of data, and this to be trained in order to obtain proper probabilities.

For the purpose of this paper, for the first part, we divided the original Europarl corpus to construct four different systems, beginning from 200,000 sentences in the smallest corpus, and increasing in steps of approximately 200,000 sentences each time up to the 4th test system with a corpus of almost 800,000 sentences.

Our Moses and cdec systems are trained in identical conditions, we used the same amounts of our considered Europarl corpus for training both the translation model and the language model.

One of the largest, freely available parallel corpus for the English-Spanish language pair is the Europarl corpus (Koehn, 2005). The domain of this corpus are general politics, economics, sciences, and technologies.

The original English-Spanish Europarl corpus consists of around 2M sentences, and around 50M words and punctuation marks, for each side. But for this experiment, our considered Europarl corpus consists of 815,000 sentences and around 21M words and punctuation marks for both English and Spanish sides.

Our tests used the English-Spanish considered Europarl parallel corpus divided into four different systems, used with both Moses and cdec toolkits.

[2]The WER is a machine translation evaluation metric, computed as the minimum number of substitution, insertion and deletion operation that have to be performed to convert the output sentence into the reference sentence

Data	Sentences	Words
English	815,000	20,595,390
Spanish	815,000	21,383,471

Table 1: Parallel corpus composition

The training part of the experiments consisted of 200,000 parallel sentences for the first experiment, 400,000 parallel sentences for the second experiment, 600,000 parallel sentences for the third experiment, and 800,000 parallel sentences for the last experiment, using in both translation directions.

The tuning part consisted of around 5000 parallel sentences of the whole corpus for using in both translation directions.

The testing part consisted of 10,000 parallel sentences with a human translation as a reference for using in both translation directions.

Systems	Training	Tuning	Testing
System 1	200,000	5,000	10,000
System 2	400,000	5,000	10,000
System 3	600,000	5,000	10,000
System 4	800,000	5,000	10,000

Table 2: Parallel corpus size for each system

3 Experiments, Results, and Evaluation

3.1 Implementation

In this paper, Moses and cdec are evaluated. We perform translation in both directions, Spanish-English and English-Spanish.

For our Moses-based experiments we set the beam size to 200, the distortion limit to 6. We limit to 20 the number of target phrases that are loaded for each source phrase, and we use the same default eight features of Moses.

Also our cdec-based experiments used the cdec implementation of the hierarchical phrase-based algorithms. Our maximum phrase length was set to 4, and maximum MIRA iterations was set to 20, with the size of N-best list at 500. The language models used are 3-gram models.

The issue of sentence alignment in the parallel corpus in use needs much attention. Sentence-aligned parallel corpora are useful for the application of machine learning to machine translation, however unfortunately it is not usual for parallel corpora to originate in this form. Several different methods are able to perform alignment. Desirable characteristics of an efficient sentence alignment method include speed, accuracy and no need for prior knowledge of the corpus or the languages in the pair.

In our experiments we used the fast-align as a simple and fast alignment tool (Dyer, Chahuneau, and Smith, 2013). All the corpora used in each test, in both the Moses and cdec experiments were aligned on sentence level, and tokenized.

3.2 Results

In this section we discuss the results we achieved, and compare Moses and cdec over our the systems that we detailed in Data Preparation section.

We trained our machine on four different systems, each with a different corpus (Table 2). Also we used the Europarl corpus for building a language model. The language model in both systems was smooth, with a modified Kneser-Ney algorithm (Pickhardt et al., 2014), and implemented in IRSTLM (Federico, Bertoldi, and Cettolo, 2008). We trained language models up to 3-grams. In our cdec tests, we used N-best list of size 500. In the final evaluation, we report the results using both BLEU and TER evaluation scores.

We start by comparing the translations yielding the best configuration generated by both cdec and Moses. In the first stage of the test, we apply Moses and cdec for the Spanish-English translation direction. As seen in (Tables 3 and 4), in system 4, we achieve the best scores. The BLEU score for Moses is **0.3144**, and for cdec is **0.3383**, and TER scores Moses at **0.5468**, and cdec at **0.5267**. The BLEU score for cdec shows a better result in comparison to Moses. The same trend is also observed in the TER score for system 4.

In the second stage of the test, we apply Moses and cdec for the English-Spanish translation direction. As seen in (Tables 5 and 6) in system 4, the TER score for Moses is **0.5331**, and for cdec is **0.5527**, and BLEU scores Moses at **0.3367** and cdec at **0.3086**. As you will observe, here Moses achieves a better score in both BLEU and TER when compared to cdec.

As we mentioned already, hierarchical phrase-based translation is based on synchronous context-free grammars (SCFG). Like classical phrase-based translation, pairs of corresponding source and target language phrases (sequences of tokens) are learned from training data. The difference is that in hierarchical models, phrases may contain gaps, and are represented by non-terminal symbols of the SCFG. If a source phrase contains a non-terminal, then the target phrase will also contain that non-terminal, and the decoder can replace the non-terminal by any source phrase and its translation respectively.

This follows the observation that hierarchical models have been shown to produce better translation results than classical phrase-based models (Chiang, 2005).

The best result report in this paper is **0.5267** TER, and **0.3383** BLEU, using the cdec toolkit trained on system 4. Moses was not able to outperform these scores, despite its ability to learn factored models. The best Moses score is **0.3367** BLEU, and **0.5331** TER. The scores indicate the hierarchical model is better than the classical model in the Spanish-English translation direction, and the classical model is better than the hierarchical one in the English-Spanish translation direction.

Tables below show the evaluation results for four different Es-En and En-Es SMT systems in terms of BLEU and TER metrics:

Translation Systems	**Moses**	**cdec**
SMT System 1	0.3085	0.3285
SMT System 2	0.3107	0.3314
SMT System 3	0.3123	0.3351
SMT System 4	**0.3144**	**0.3383**

Table 3: BLEU scores Es-En cdec vs. Moses

Translation Systems	**Moses**	**cdec**
SMT System 1	0.5515	0.5327
SMT System 2	0.5507	0.5312
SMT System 3	0.5489	0.5292
SMT System 4	**0.5468**	**0.5267**

Table 4: TER scores Es-En cdec vs. Moses

Translation Systems	Moses	cdec
SMT System 1	0.3281	0.3019
SMT System 2	0.3312	0.3033
SMT System 3	0.3338	0.3051
SMT System 4	**0.3367**	**0.3086**

Table 5: BLEU scores En-Es cdec vs. Moses

Translation Systems	Moses	cdec
SMT System 1	0.5416	0.5595
SMT System 2	0.5381	0.5567
SMT System 3	0.5356	0.5546
SMT System 4	**0.5331**	**0.5527**

Table 6: TER scores En-Es cdec vs. Moses

4 Discussion

In order to compare the classical phrase-based translation model with the hierarchical phrase-based translation model we decided to use Moses, which is a phrase-based translation toolkit, and cdec, which is the hierarchical phrase-based translation toolkit, for both Spanish-English and English-Spanish translation directions.

Basically, classical phrase-based translation models and hierarchical phrase-based translation models have different strengths and weaknesses. Classical models translate better with local reordering and hierarchical models translate better with slightly longer range reordering. Classical models do tend to get higher BLEU scores and lower TER scores while hierarchical models often do better in human evaluation for language pairs that have more long distance reordering.

In general, when we add the training data to each system for evaluating the performance of the translation system, the BLEU score has to increase, and the TER score has to decrease. In order to compare two different translation systems, after adding the training data, if the new BLEU score increased than the previous system, and the new TER score decreased than the previous system, it means that, the new considered system is working well. So adding more training data to the translation system for both translation directions helps significantly improve the BLEU and the TER scores.

5 Conclusion and Future Work

In this paper we used four different data collections for comparing the performance of the Spanish-English and the English-Spanish statistical machine translation systems, in order to find the best translation model for each translation direction. We showed different behavior of the classical phrase-based translation model and the hierarchical phrase-based translation model on four different data sets for Spanish/English language pair statistical machine translation and we observed several results.

In our experiments after adding the training data, both the BLEU and the TER scores improved when we trained with cdec in the Spanish-English translation direction, whereas Moses had a better performance in the English-Spanish translation direction. So cdec had a better performance in the Spanish-English direction and Moses had a better performance in the English-Spanish direction. It means the hierarchical phrase-based translation model has a better performance for Spanish-English translation direction and the classical phrase-based translation model has a better performance in back translation direction.

In our future work we want to explore problems with existing data sets, the issue of morphology and its relation to output quality by combining those models together. Also we want to use the hierarchical phrase-based translation model, in order to train a statistical machine translation system for some language pairs without enough parallel data such as Spanish-Persian translation direction and back translation using pivot language technique through cdec platform. cdec as a good hierarchical decoder can capture word order even better than Moses. Its results tend to be always slightly better in the Spanish to English translation direction.

Acknowledgements

This article is supported by the Catalan Government Grant Agency, Ref. 2014SGR027.

References

Chiang, D. 2005. A hierarchical phrase-based model for statistical machine translation. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *ACL*. The Association for Computer Linguistics.

Chiang, D. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33.

Dyer, C., V. Chahuneau, and N. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of NAACL-HLT 2013*, pages 644–648, Atlanta, Georgia.

Dyer, C., A. Lopez, J. Ganitkevitch, J. Weese, H. Setiawan, F. Ture, V. Eidelman, P. Blunsom, and P. Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *In Proceedings of ACL System Demonstrations*.

Federico, M., N. Bertoldi, and M. Cettolo. 2008. Irstlm: an open source toolkit for handling large scale language models. In *Interspeech*, pages 1618–1621. ISCA.

Hutchins, W. 1995. Machine translation: A brief history. In *concise history of the language sciences: from the sumerians to the cognitivists, pergamon*, pages 431–445. Press.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Koehn, P., 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 *ACL*. The Association for Computer Linguistics.

Koehn, P., F.J. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ. Association for Computational Linguistics, Association for Computational Linguistics.

Lopez, A. 2008. Statistical machine translation. *ACM Comput. Surv.*, 40(3):8:1–8:49, August.

Marcu, D. and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *In Proceedings of EMNLP*, pages 133–139.

Och, F.J. and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Pickhardt, R., T. Gottron, M. Korner, P. Wagner, T. Speicher, and S. Staab. 2014. A generalized language model as the combination of skipped n-grams and modified kneser ney smoothing. In *ACL'14: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Ture, F. and J. Lin. 2013. Flat vs. hierarchical phrase-based translation models for cross-language information retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 813–816, New York, NY, USA. ACM.

White, B. and E. Cambria. 2014. Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9:2.