Calculation of phrase probabilities for Statistical Machine Translation by using belief functions

Christophe Servan, Simon Petitrenaud

Calculation of phrase probabilities for Statistical Machine Translation by using belief functions

Christophe SERVAN *Simon PETITRENAUD*
University of Le Mans, Avenue Laënnec 72085 Le Mans Cedex 9, France
first-name.last-name@lium.univ-lemans.fr

ABSTRACT
In this paper, we consider a specific part of statistical machine translation: feature estimation for the translation model. The classical way to estimate these features is based on relative frequencies. In this new approach, we propose to use the concept of belief masses to estimate the phrase translation probabilities. The Belief Function theory has proven to be suitable and adapted for dealing with uncertainties in many domains. We have performed a series of experiments to translate from English into French and from Arabic into English showing that our approach performs, at least as well as and at times better than, the classical approach.

KEYWORDS: Belief function, Statistical Machine Translation.

Proceedings of COLING 2012: Posters, pages 1101–1110,
COLING 2012, Mumbai, December 2012.

1 Introduction

In statistical machine translation (SMT), there have been many works on smoothing translation model probabilities (Foster et al., 2006; Kuhn et al., 2010), but few work on feature estimation. (Chiang et al., 2009) proposed to add new features outside the translation model, but to the best of our knowledge there is few research on a different way to estimate the features of the translation model (TM). De Nero and Moore (DeNero et al., 2006; Moore and Quirk, 2007) proposed some approaches that did not improve the translation. More recent works based on a smooth Maximum-Likelihood estimate (Sima'an and Mylonakis, 2008) give better results. As we consider the popular phrase-based approach, the TM corresponds to the phrase table in this paper.

The phrase table is basically a list of possible translations and their probabilities for a given source phrase. Each line or event of a phrase table is composed of a source and a target language phrase pair. The events are supposed to be independent from each other. Phrase tables may contain many features like phrase translation and lexical probabilities. In order to estimate these probabilities, SMT uses very large corpora called *bitexts*, which are composed of sentences translated from a source language into a target language. For each sentence, the words of both languages are aligned according to the translation.

In the classical approach, the estimation of the probabilities is performed by the use of simple count functions, based on relative frequencies. But it is possible to use other concepts to estimate the features. In particular, many authors showed that the Dempster-Shafer theory (or Belief Function theory) allows a more flexible representation of uncertainty than a probability model (Smets, 1988; Cobb and Shenoy, 2006). For example, probabilities do not really take into account the conflict between different translation hypotheses, especially in the case of rare examples, or the global confidence in the translations. The belief function theory, as an alternative to the probability theory, can take this into account. In this paper, we present an original way to estimate the feature associated with a set of phrase pairs with the use of belief functions.

This paper presents our first studies and results obtained with this new approach. Firstly, we briefly recall the theory of SMT. In Section 3, we present our approach based on belief functions. Then, we propose several experiments in order to show the effectiveness of our approach. At last we conclude this paper and present some perspectives.

2 Background

2.1 General model for statistical machine translation

Let us assume that we are given a source sentence s to be translated into different target sentences $t_i \in T_s$, where T_s is the set of all observed translations of s in the phrase table. The statistical machine translation (SMT) model uses a set of n feature functions $f_k, k = 1 \ldots, n$, depending on the source and target word sequences, in order to estimate the best translation. Typical feature functions include the translation and distortion model, a language model on the target language and various penalties. Among all possible target sentences, the sentence is chosen as follows:

$$t^* = \arg\max_{t_i \in T_s} \log \left(\prod_{i=1}^{n} f_k(t_i, s)^{\lambda_k} \right), \quad (1)$$

where each parameter λ_k is a coefficient to weight the feature function f_k (Och, 2003). These weights are usually optimized so as to maximize the translation performance on some devel-

opment dataset. The work presented in this paper focuses on features used to estimate the translation model.

2.2 Feature estimation in statistical machine translation

In the popular Moses toolkit (Koehn et al., 2007), the phrase table contains five features (Koehn, 2010): the phrase translation features and the lexical weighting for both translation directions, and the phrase penalty. The phrase translation features are usually estimated using relative frequency; the lexical weights are estimated by using the word-based IBM Model 1 of each phrase pair. At last, the phrase penalty depends on phrase length. This feature is set by the user to the same value ρ for each phrase. If $\rho > e$, longer phrases will be preferred over shorter ones. Conversely, if $\rho < e$, shorter phrases will be preferred.

Source language (s) - fr	Target language (t) - en
...	...
étant donné un	given a
étant donné un	starting from an
étant donné	given
étant donné	given
étant donné	starting from
étant donné	starting
...	...

Table 1: Example of phrase pairs extracted from a bitext.

Table 1 gives an example of phrase pairs extracted from a bitext and a small part of the corresponding phrase table is presented in Table 2. In this example, the classical estimation of the feature of the phrase translation pair "*starting*" given "*étant donné*" is equal to 0.25 and the probability of "*given*" given "*étant donné*" is equal to 0.5. The inverse phrase translation probability is estimated in the same way.

source (s) - fr	seg cib (t) - en	$p(t\|s)$	$lex(t\|s)$	$p(s\|t)$	$lex(s\|t)$
...	...				
étant donné	given	**0.5**	0.060147	**0.333333**	0.306373
étant donné	starting	**0.25**	7.15882e-06	**0.333333**	5.19278e-05
étant donné	starting from	**0.25**	7.15882e-06	**0.333333**	0.0277778
...	...				

Table 2: Extract of a translation table with parameters.

This classical way to estimate the phrase translation probability may have some drawbacks. When some unique phrase translation pair occurs many times, like the pair "*la maison blanche*|*the white house*", the phrase translation probability estimation is equal to 1. But in other situations, occurrences are very rare and ambiguous at the same time. For example, let us assume that for the French word "$dents$" (which should be translated as "$teeth$" in English), two contradictory pairs are available in the phrase table: "$dent|teeth$" and "$dents|jaws$". These events may both have a probability estimation equal to 1 because they occur only once.

Even though the estimation of the inverse phrase translation pair may balance this problem, if the event is observed only once in either translation direction, the inverse estimation is useless. The goal of this work is to propose a new way to estimate the translation features in both translation directions. Fortunately, thanks to alternative theories to the probability theory, it is possible to improve these estimations. One of these theories is particularly suited to deal with uncertainties: the theory of belief functions, which has been developed for thirty years. This theory has been successfully applied in several domains such as speaker identification

$m_s(\text{starting}) = p(\text{starting}\|\text{étant donné}) * \overline{p(\text{given}\|\text{étant donné})} * \overline{p(\text{starting from}\|\text{étant donné})}$
$m_s(\text{starting}) = 0.09375$
$m_s(\text{starting from}) = 0.09375$
$m_s(\text{given}) = 0.28125$

Table 3: Example of the estimation of some phrase pair features with the TBM Theory (s ="étant donné")

(Petitrenaud et al., 2010) or classification (Elouedi et al., 2000). In this paper, we adapt some concepts of this theory to our feature estimation problem.

3 Belief functions for SMT

In this section, we briefly present some notions of the belief function theory (Shafer, 1976; Smets and Kennes, 1994) and we apply it to the problem of feature estimation. In this article, we adopt the point of view proposed by Smets: the Transferable Belief Model (TBM) (Smets and Kennes, 1994). The aim of this model is to determine the belief concerning different propositions, from some available information.

3.1 Belief function theory

Let Ω be a finite set, called frame of discernment of the experience. The representation of uncertainty is made by the means of the concept of belief function, defined as a function m from 2^Ω to $[0,1]$ as $\sum_{A\subseteq\Omega} m(A) = 1$. The quantity $m(A)$ represents the belief exactly allowed to proposition A. The subsets A of Ω such that $m(A) > 0$ are called focal elements of m. In the very particular case when Ω is the only focal element (i.e. $m(\Omega) = 1$ and $\forall A \neq \Omega, m(A) = 0$), the belief function expresses a total lack of information on the frame of discernment. This is one of the essential differences with the probability theory. In general, the total absence of information would be represented by a uniform distribution on Ω in probability theory. One of the most important operations in the TBM is the procedure of aggregating information to combine several belief functions defined in a same frame of discernment (Smets and Kennes, 1994). In particular, the combination of two belief functions m_1 and m_2 independently defined on Ω using the conjunctive binary operator $\cap$, denoted as $m' = m_1 \cap m_2$, is defined as (Smets and Kennes, 1994) :

$$\forall A \subseteq \Omega, m'(A) = \sum_{B\cap C=A} m_1(B) * m_2(C). \tag{2}$$

Note that this combination operation may produce a non-null mass on the empty set $\emptyset$. The quantity $m'(\emptyset)$ represents the mass that cannot be allocated to any proposition of Ω. In this case, $m'(\emptyset)$ also measures the conflict between the belief functions m_1 an m_2. Since the operator $\cap$ is commutative and associative, it is easy to define the combination of n functions $m_1, \ldots, m_n$ on Ω by:

$$m = m_1 \cap \ldots \cap m_n = \bigcap_{i=1}^{n} m_i, \tag{3}$$

with $m(A) = \sum_{A_1\cap\ldots\cap A_n=A} \prod_{i=1}^{n} m_i(A_i), \quad \forall A \subseteq \Omega$. Finally the function m captures the global information concerning the experience.

3.2 Belief functions as features for the translation model

Here, we propose to use the TBM to estimate the phrase translation features. First, for a given source s, each target $t_i \in T_s$ gives a piece of information for the translation and can be described by a belief function m_s^i, such as:

$$\begin{cases} m_s^i(\{t_i\}) = p(t_i|s) \\ m_s^i(T_s) = \overline{p(t_i|s)}, \end{cases} \tag{4}$$

where $\overline{p(t_i|s)} = 1 - p(t_i|s)$. The belief function m_s^i has only two focal elements: t_i, and T_s. The mass on T_s expresses a confidence degree for this piece of information. We combine the information defined by all the translation hypotheses concerning s, thanks to the conjunctive obtained by the following straightforward formula:

$$m_s = \bigcap_{t \in T_s} m_s^i. \tag{5}$$

The resulting mass concerning t_i is obtained by the following formula (cf. Equation 3):

$$m_s(\{t_i\}) = p(t_i|s) * \prod_{t_k \in T_s \setminus \{t_i\}} \overline{p(t_k|s)}. \tag{6}$$

Note that $\sum_{t_i \in T_s} m_s(\{t_i\}) = 1 - m(T_s) - m(\emptyset) < 1$ generally. The mass $m(T_s)$ and $m(\emptyset)$ can be interpreted respectively as the general ignorance degree and the level of information conflicting concerning the translation of s. Then we obtain our feature estimation defined in Equation 1 by: $f(t_i, s_j) = m_{s_j}(\{t_i\})$. In the same way, we obtain the inverse feature estimation by the following equation:

$$m_t^I(\{s_j\}) = p(s_j|t) * \prod_{s_k \in S_t \setminus \{s_j\}} \overline{p(s_k|t)}, \tag{7}$$

where S_t is the set of possible sources for target t. If we apply these formulas to the example presented in Tables 1 and 2, the new estimation of the features associated with the phrase translation pairs are computed in Table 3. Note that if $p(t_i|s) = 1$, the belief masses for the other hypotheses become zero (cf. Equation 6). The belief mass indicated in this equation may be modified as follows: $m_s(\{t_i\}) = \frac{1}{1 + \frac{1}{|s|}}$, where $|s|$ denotes the count of s. Thus, $m_s(t_i) < 1$ but $m_s(t_i)$ tends to 1 when the information concerning s increases. Finally, the optimized target sentences are obtained by Equation 1.

4 Experimental design

As presented before, the new approach with the TBM is applied only on the phrase translation features. Therefore, in all our experiments, we have removed the lexical features from the phrase tables of all the translation models. We performed several experiments on various language pairs with various kinds of corpus kind described in the next part. The metrics used are the BLEU score (Papineni et al., 2002) and the TER metric (Snover et al., 2006).

4.1 Data

Several data sets were used in our experiments, with various language pairs and various kind of data (*e.g.* news, scientific papers). A complete description of all the corpora presented in this

part is shown in Table 4. The framework used in the evaluation of the WMT task contains a set of several corpora. The corpora used in our experiments are described in Table 4. The training corpora used are Europarl 7 (eparl7), News-commentary 7 (nc7).

Task COSMAT	corpus	AbsTrain		AbsDev		AbsTest		
	language	fr	en	fr	en	fr	en	
	# of sentences	5141		1083		1102		
	# of words	135K	120K	28K	25K	28K	25K	
Task WMT	corpus	nc7+eparl7		nwtst2010		nwtst2011		
	language	fr	en	fr	en	fr	en	
	# of sentences	2M		2489		3003		
	# of words	65,7M	59M	62k	70k	75k	84,5k	
Task Ar-En	corpus	train		nist09-nw		nist08-nw		
	language	ar	en	ar	en	ar	en	
	# of sentences	184K		586		813		
	# of words	4.8M	5M	23K	23K	29K	28K	

Table 4: Description of the bitexts and development (or tuning) and test corpora.

The corpus from the French ANR project COSMAT[1] is composed of a collection of abstracts of PhD Theses in both French and English (Lambert et al., 2012). These abstracts have been classified according to several topics. In our experiments, we selected only the topic of computer science.

The last set of data concerns the translation of Arabic news into English. Following the GALE program, the DARPA launched a new 5 year language technology program called Broad Operational Language Translation program (BOLT). The goal of this project is to create a technology capable of translating multiple foreign languages in all genres, to retrieve information from the translated material, and to enable a bilingual communication via speech or text from Arabic and Mandarin into English. We used as the development corpus the news part of the NIST 2009 evaluation (*nist09-nw*) and as the test corpus the NIST 2008 evaluation (*nist08-nw*). The system ensue from this corpus could be used in the NIST evaluation.

4.2 Stability test

In order to have a more reliable precision in our experiments, we performed several optimizations with random initialisation toward the BLEU score for each experiment (Clark et al., 2011). Following this method, three runs of Minimum Error Rate Training (MERT) (Och, 2003) and Margin Infused Relaxed Algorithm (MIRA) (Chiang et al., 2008, 2009) were made. Then, the result is an average of these three runs and the standard deviation is given between parenthesis next to the scores. Both optimization approaches were used to observe how these features are influenced by the process.

4.3 Results

Tables 5, 6 and 8 show the results obtained with the classical approach and with our approach based on the TBM theory. The Brevity Penalty is about 0.99 (0.01) for the two approaches in each experiment.

First, we compare the classical ($Proba.$) and the TBM ($Belief$) approaches with the tuning thought MERT. According to the French-English WMT (Table 5), we can observe a slight improvement of the BLEU score when the translation is from French into English. The $Belief$

[1] http://www.cosmat.fr

optimization process		MERT		MIRA	
corpus	approach	BLEU	TER	BLEU	TER
Translation direction: fr↦en					
nwtst2010	Proba.	27.42 (0.04)	54.50 (0.03)	27.22 (0.04)	54.48 (0.04)
(Dev)	Belief	27.47 (0.06)	54.46 (0.03)	27.21 (0.04)	54.53 (0.03)
	Proba.+Belief	27.43 (0.05)	54.61 (0.09)	**27.34** (0.01)	**54.54** (0.05)
nwtst2011	Proba.	27.69 (0.07)	53.85 (0.04)	27.73 (0.04)	53.75 (0.05)
(Test)	Belief	27.72 (0.03)	53.87 (0.06)	27.79 (0.03)	53.74 (0.04)
	Proba.+Belief	27.59 (0.11)	53.95 (0.12)	27.79 (0.04)	53.80 (0.06)
Translation direction: en↦fr					
nwtst2010	Proba.	26.55 (0.05)	58.67 (0.03)	26.41 (0.04)	58.44 (0.21)
(Dev)	Belief	26.51 (0.09)	58.88 (0.06)	26.42 (0.08)	58.58 (0.11)
	Proba.+Belief	**26.64 (0.05)**	**58.66** (0.20)	**26.53** (0.06)	**58.42** (0.07)
nwtst2011	Proba.	28.29 (0.31)	56.74 (0.19)	28.59 (0.05)	56.18 (0.15)
(Test)	Belief	28.39 (0.07)	56.88 (0.06)	28.72 (0.03)	56.28 (0.05)
	Proba.+Belief	**28.47** (0.06)	**56.72** (0.21)	**28.74** (0.06)	**56.06** (0.12)

Table 5: BLEU and TER scores obtained on the WMT task.

approach can be considered as efficient as the classical one. The performance gain is more visible when the translation is from English to French and can reach about 0.18 BLEU point on the test corpus. In Table 6, the experiment shows a decrease of the two scores (respectively 0.1 and 0.15) on the tuning corpus. Contrary to the tuning corpus, an improvement of 0.1 BLEU point and 0.08 point of TER is visible on the test corpus. In the last experiment proposed on the COSMAT corpus (Table 8), we can observe a decrease only on the tuning corpus when we translate French into English. On the test corpus, the improvement is about 0.1 point on both BLEU and TER score. When the translation is from English into French, the increase is about 0.04 BLEU point on both development and test corpus.

optimization process		MERT		MIRA	
corpus	approach	BLEU	TER	BLEU	TER
nist09-nw	Proba.	**33.81** (0.14)	**47.27** (0.13)	34.04 (0.07)	48.51 (0.19)
(Dev)	Belief	33.71 (0.07)	47.42 (0.03)	34.11 (0.03)	48.08 (0.34)
	Proba.+Belief	33.74 (0.14)	47.56 (0.33)	**34.30** (0.06)	**48.14** (0.12)
nist08-nw	Proba.	26.99 (0.09)	57.37 (0.18)	26.51 (0.09)	58.54 (0.25)
(Test)	Belief	**27.10** (0.08)	**57.29** (0.10)	**26.83** (0.14)	**57.71** (0.32)
	Proba.+Belief	26.90 (0.09)	57.65 (0.28)	26.73 (0.13)	58.24 (0.36)

Table 6: Results for the translation from Arabic into English.

Regarding the results, the new approach is at least as efficient as the classical one. When we look at the entropy of each phrase table, and the various translations produced by the systems, we can observe better translations in the two approaches like in the example given in Table 7. This example shows us how the combination takes the best of the two approaches.

Source	ils permettent la réutilisation du code de gestion de la duplication et de la cohérence .
Reference	they allow reuse of replication and consistency management code .
Proba.	they allow the reuse of the code of management and replication consistency .
Belief	they allow the reuse of the code of replication and the consistency .
Proba.+Belief	they allow the reuse of the code of replication management and consistency .

Table 7: Example of translation from French into English.

This has lead us to combine the two approaches ($Proba. + Belief$) in all experiments. The increase is visible in the WMT task on the translation of English into French about 0.2 BLEU point on the tuning corpus and 0.15 BLEU point on the test corpus. With the COSMAT experiment in table 8, in both translation directions when using the combined approaches, we observe an increase of the BLEU score.

These experiments show us this novel approach can be considered as comparable to the classical

optimization process		MERT		MIRA	
corpus	approach	BLEU	TER	BLEU	TER
Translation direction: fr→en					
AbsDev	Proba.	35.83 (0.03)	47.64 (0.20)	35.54 (0.03)	47.98 (0.05)
	Belief	35.82 (0.06)	47.89 (0.04)	35.66 (0.04)	**47.71** (0.03)
	Proba.+Belief	**35.93 (0.06)**	47.80 (0.05)	**35.72** (0.04)	47.72 (0.04)
AbsTest	Proba.	43.02 (0.11)	42.73 (0.17)	**43.00** (0.01)	**42.61** (0.03)
	Belief	43.13 (0.09)	42.66 (0.05)	42.81 (0.10)	42.64 (0.00)
	Proba.+Belief	**43.27** (0.17)	42.62 (0.12)	42.95 (0.04)	42.63 (0.01)
Translation direction: en→fr					
AbsDev	Proba.	41.95 (0.20)	45.68 (0.39)	42.06 (0.01)	46.23 (0.03)
	Belief	41.99 (0.18)	46.25 (0.20)	42.06 (0.07)	46.17 (0.06)
	Proba.+Belief	42.12 (0.10)	46.08 (0.10)	**42.18** (0.12)	**46.03** (0.04)
AbsTest	Proba.	33.33 (0.09)	**52.22** (0.82)	33.15 (0.05)	53.04 (0.05)
	Belief	33.37 (0.08)	52.83 (0.28)	33.26 (0.08)	52.92 (0.03)
	Proba.+Belief	**33.56** (0.10)	52.72 (0.20)	**33.45** (0.06)	**52.81** (0.08)

Table 8: Results obtains on the COSMAT Task.

approach and can be more efficient under certain conditions. But for all the systems tuned with MERT, there is sometimes a high standard deviation of about 0.2 BLEU point and 0.3 TER point. Recent MIRA experiments (Cherry and Foster, 2012) show a lower deviation of the score and a better robustness than MERT. All the experiments were rerun with MIRA in order to observe a smaller deviation and a better precision in our experiments; this is the second set of experiments shown in the various tables.

The new set of experiment shows an improvement especially for the WMT results (Table 5): the combination of the two approaches obtains 0.15 BLEU point and 0.1 TER point more than the classical approach. The improvement, visible in Table 6, reach respectively 0.26 and 0.2 BLEU point and 0.37 and 0.3 TER point on the tuning corpus and the test corpus. At last, in the COSMAT experiment, a decrease is visible when we translate French into English on the test corpus of 0.05 BLEU point. But when we translate English into French, the increase can reach 0.3 BLEU point and 0.23 TER point on the test corpus. It seems the MIRA process give a better advantage to the combination of the two approaches over the classical approach. We can also observe better results comparing the two optimizations in the WMT task when we translate from English into French.

Conclusions and Further Work

In this paper, we presented our first results on the application of the Transferable Belief Model to Statistical Machine Translation. The approach was used to estimate only the phrase translation pair features. The results obtained on the different experiments lead us to combine the new and the classical approaches. The score of the translations obtained with this combination is improved, and this also leads to better translation quality according our experiments. This combination of approaches encourages us to work further. For example, this new approach could be applied as a secondary phase-table in order to rescore the first one. This rescoring could be done during the decoding process on the graph of hypothesis or on the n-best translations output as proposed in several works on language model adaptation and rescoring (Bacchiani and Roark, 2003; Schwenk et al., 2006; Bulyko et al., 2007).

Acknowledgments

This work has been partially funded by the European Union under the EuroMatrixPlus project ICT-2007.2.2-FP7-231720, the French government under the ANR project COSMAT ANR-09-CORD-004 and the DARPA Bolt project.

References

Bacchiani, M. and Roark, B. (2003). Unsupervised language model adaptation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, Hong Kong, China.

Bulyko, I., Matsoukas, S., Schwartz, R., Nguyen, L., and Makhoul, J. (2007). Language model adaptation in machine translation from speech. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, Honolulu, Hawaii, USA.

Cherry, C. and Foster, G. (2012). Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada. Association for Computational Linguistics.

Chiang, D., Knight, K., and Wang, W. (2009). 11,001 new features for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 218–226, Boulder, Colorado, USA. Association for Computational Linguistics.

Chiang, D., Marton, Y., and Resnik, P. (2008). Online large-margin training of syntactic and structural translation features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 224–233, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA. Association for Computational Linguistics.

Cobb, B. R. and Shenoy, P. P. (2006). A comparison of methods for transforming belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):255–266.

DeNero, J., Gillick, D., Zhang, J., and Klein, D. (2006). Why generative phrase models underperform surface heuristics. In *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, pages 31–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Elouedi, Z., Mellouli, K., and Smets, P. (2000). Classification with belief decision trees. In *Proceedings of the 9th International Conference on Artificial Intelligence : Methodology, Systems, Architectures.*, Varna, Bulgaria. AIMSA 2000, Springer Lecture Notes on Articial Intelligence.

Foster, G., Kuhn, R., and Johnson, H. (2006). Phrasetable smoothing for statistical machine translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 53–61, Sydney, Australia. Association for Computational Linguistics.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Kuhn, R., Chen, B., Foster, G., and Stratford, E. (2010). Phrase clustering for smoothing tm probabilities: or, how to extract paraphrases from phrase tables. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 608–616, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lambert, P., Senellart, J., Romary, L., Schwenk, H., Zipser, F., Lopez, P., and Blain, F. (2012). Collaborative machine translation service for scientific texts. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 11–15, Avignon, France. Association for Computational Linguistics.

Moore, R. C. and Quirk, C. (2007). An iteratively-trained segmentation-free phrase translation model for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 112–119, Stroudsburg, PA, USA. Association for Computational Linguistics.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Petitrenaud, S., Jousse, V., Meignier, S., and Estève, Y. (2010). Automatic named identification of speakers using belief functions. In *Information Processing and Management of Uncertainty (IPMU'10)*, Dortmund, Germany.

Schwenk, H., Déchelotte, D., and Gauvain, J.-L. (2006). Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 723–730.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Sima'an, K. and Mylonakis, M. (2008). Better statistical estimation can benefit all phrases in phrase-based statistical machine translation. In *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT) 2008*, pages 237–240, Goa, India.

Smets, P. (1988). Belief functions versus probability functions. In Bouchon, B., Saitta, L., and Yager, R., editors, *Uncertainty and Intelligent Systems*, volume 313 of *Lecture Notes in Computer Science*, pages 17–24. Springer Berlin / Heidelberg.

Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, 66:191–234.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7^{th} Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA.