The UPC TweetMT participation: Translating Formal Tweets Using Context Information*

Participación de la UPC en TweetMT: Traducción de tweets formales usando información de contexto

Eva Martínez Garcia
Cristina España-Bonet
TALP Research Center
Universitat Politècnica de Catalunya
Jordi Girona, 1-3, 08034 Barcelona, Spain
{emartinez,cristinae}@cs.upc.edu

Lluís Màrquez
Qatar Computing Research Institute
Qatar Foundation
Tornado Tower, Floor 10,
P.O. Box 5825, Doha, Qatar
lmarquez@qf.org.qa

Resumen: En este artículo describimos los sistemas con los que participamos en la tarea compartida TweetMT. Desarrollamos dos sistemas para el par de idiomas castellano–catalán: un traductor estadístico diseñado a nivel de frase y un sistema sensible al contexto aplicado a tweets. En el segundo caso definimos el "contexto" de un tweet como los tweets producidos por un mismo usuario durante un día. Estudiamos el impacto de este tipo de información en las traducciones finales cuando se usa un traductor a nivel de documento. Una variante de este sistema incluye modelos semánticos adicionales.
Palabras clave: Traducción automática, Twitter, Traducción sensible al contexto

Abstract: In this paper, we describe the UPC systems that participated in the TweetMT shared task. We developed two main systems that were applied to the Spanish–Catalan language pair: a state-of-the-art phrase-based statistical machine translation system and a context-aware system. In the second approach, we define the "context" for a tweet as the tweets of a user produced in the same day, and also, we study the impact of this kind of information in the final translations when using a document-level decoder. A variant of this approach considers also semantic information from bilingual embeddings.
Keywords: Machine Translation, Twitter, Context Aware Translation

1 Introduction

Twitter is a very popular social network. This microblogging service allows users to share a huge amount of information in a quick way. Usually, Twitter users produce monolingual content (34% in English and 12% in Spanish for example[1]). However, Twitter is a multilingual communication environment. There are many users from different nationalities posting messages in their own language. So, to ease the spread of the information, it would be useful to post messages in several languages simultaneously. One option to create multilingual tweets is by crowdsourcing manual translations. Meedan is a non-profit organization[2] which uses this resource to share news between the Arabic and the English speaking communities. Another example, although in this case applied to SMS, is the work done with a crowdsourced translation during the earthquake in Haiti in 2010 (Munro, 2010). They allowed the Haitian Kreyol and French-speaking communities of volunteers to translate texts into English, categorize and geolocate the messages in real–time in order to help the primary emergency responders. There are also a few works applying machine translation to tweets and the interest in the topic is growing over the years. In (Gotti, Langlais, and Farzindar, 2013), the application of statistical machine translation (SMT) systems to translate tweets from the Canadian Government Agen-

* This research has been partially funded by the TACARDI project (TIN2012-38523-C02) of the Spanish Ministerio de Economía y Competitividad (MEC). The authors thank Iñaki Alegría and Gorka Labaka for providing monolingual corpora.

[1] http://www.technologyreview.com/graphiti/522376/the-many-tongues-of-twitter/

[2] http://news.meedan.net

cies is studied. The authors in (Jehl, Hieber, and Riezler, 2012) describe a system that does not rely on parallel data. In contrast, they try to find similar tweets in the target language in order to train a standard phrase-based SMT pipeline.

All previous papers describe some common problems when trying to translate tweets or short messages. The first and most usual obstacle is the colloquial language used in the messages, closely followed by the writing errors. To address these phenomena, it is necessary to apply a normalization step previous to translation. Also, the Twitter 140-character constraint is hard to maintain in a translation, so an effort must be made to generate legal length tweets. Another very common problem is handling the hashtags. It is not clear whether they have to be translated or not, as well as their position in the sentence.

TweetMT is a shared task with the aim of translating formal tweets[3]. These are messages usually tweeted by institutions and are well written and with no use of colloquial vocabulary.

In this paper we introduce the two systems presented to the competition for the Spanish–Catalan language pair and also some of the improvements made after the submission deadline. First, we present a state-of-the-art SMT system adapted and tuned using Twitter messages. Second, we present a system that looks at the context information of a tweet to improve its translation. This second system uses a document-level decoder to take into account the context and it can be combined with bilingual distributed vector models, which allow to consider additional semantic information.

This paper is organized as follows. We describe the developed systems in Section 2 and analyze the obtained results in Section 3. Finally, we present some discussions and guidelines for future work in Section 4.

2 System Description

This section describes the corpora used for training the systems (both general corpora and corpora of tweets) and their processing as a common resource for the two main translation engines.

[3]All the information and resources related to the TweetMT2015 shared task are available at: `http://komunitatea.elhuyar.org/tweetmt/`

2.1 Data

As parallel corpus we use the Spanish–Catalan corpus *El periódico* which is a collection of news with 2,478,130 aligned sentences and available at the ELRA catalogue[4]. The shared-task organization released 4,000 parallel tweets for development and 2,000 for testing.

In order to adapt the systems to the Twitter genre, we also gather a collection of monolingual tweets. The Catalan corpus of tweets was collected using the Twitter API during the period going from 13th March 2015 to 8th May 2015. We selected 65 users with accounts mainly coming from Catalan institutions, sport clubs or newspapers. This way, we expect users to post mostly using a formal language. It is worthy to notice that there is some overlap between the users that we selected and the ones considered in the TweetMT corpora. In particular, we used some tweets from *museupicasso*, *Liceu_cat*, *Penya1930* and *RCDEspanyol*. Since the TweetMT test and development data were collected in 2013–2014 and our monolingual tweet corpora in 2015, there is no overlap between the training data and the tweets corpora delivered for the task. 90,744 tweets in Catalan were obtained with this methodology. A similar corpus in Spanish was already available as a resource of the Tweet Normalization Workshop (Alegria et al., 2014)[5]. In this case, 227,199 were collected only in two days, 1st and 2nd of April 2013.

We also use standard monolingual corpus to build larger language models. On the one hand, the corpora available in Catalan in the Opus site[6] are selected (4.8M sentences). On the other hand, the corpora provided for the WMT13 Quality Estimation Task[7] are used for Spanish (53.8M sentences).

We pre-processed the development dataset and the monolingual corpora of tweets in order to make them similar to the format of the

[4]`http://catalog.elra.info/product_info.php?products_id=1122`

[5]`http://komunitatea.elhuyar.org/tweet-norm`

[6]`http://opus.lingfil.uu.se/`, Corpora: DOGC, KDE4, OpenSubtitles 2012 and 2013, Ubuntu and Tatoeba corpora (Tiedemann, 2012; Tiedemann, 2009) (66.5M words).

[7]`http://statmt.org/wmt13/quality-estimation-task.html`, Corpora: Europarl corpus v7; United Nations; NewsCommentary 2007, 2008, 2009 and 2010; AFP, APW and Xinhua (1.59G words).

test set. That includes changing every URL in the data for the URLURLURL label and substituting every username by the IDIDID label. We decided not to translate hashtags due to their difficulty and because we observed that, in the development set, approximately two thirds of them remained untranslated. In order to maintain the hashtag information, we replace every hashtag in a tweet by a Hn label, where n is the number of hashtag, and we maintain a record file where the hashtags that appear in a tweet are stored. This strategy allows us to generalize the translation for every hashtag and eases the replacement by the corresponding original value before building the final translation. The position of the hashtags in our systems is determined by the position assigned to the corresponding labels by the decoder.

2.2 Basic SMT System

Our basic approach is a state-of-the-art phrase-based SMT system based on the Moses decoder (Koehn et al., 2007) and GIZA++ (Och and Ney, 2003). We trained the system using the *El periódico* Spanish–Catalan parallel corpus.

Language models were built using the SRILM toolkit (Stolcke, 2002). The Spanish general language model is an interpolation of several 5-gram language models with interpolated Kneser-Ney discounting as given by (Specia et al., 2013)[8]. The Catalan 5-gram language model has been built with the same features on the general Catalan monolingual corpus explained above. In order to adapt the Moses system to the Twitter genre, we introduced a second language model trained using only the tweet corpora described in the previous subsection. The Moses decoder uses both language models as feature functions.

Finally the system is tuned with MERT (Och, 2003) against the BLEU measure (Papineni et al., 2002) on the tweets of the development set.

2.3 Context-Aware SMT System

A current limitation of standard SMT systems is the fact that they translate sentences one after the other without using the information given by the surrounding ones. This problem can be even more pronounced in short sentences such as tweets where the number of content words is very small (17 words/tweet in average in the develoment set for both languages).

In order to alleviate this limitation, we use a document-level decoder that takes as a translation unit a whole document. In our case, one has to define first what a document is. After analyzing the development data, we decided to define the *context* of a tweet as the surrounding tweets posted by the same user during the same day. In cases where this number was less than 30, we put together the tweets posted during consecutive days until reaching the threshold of, at least, these 30 tweets. In that way, we expect to obtain collections of tweets –a document– that are closely related, since they come from the same source and they have been produced in a short lapse of time. Notice that this way of choosing the related tweets does not reflect a real scenario on Twitter where only the previous tweets from a particular user are available. However, in an offline scenario, considering past and future context will caracterize better the domain of the messages. We left as future work to compare the differences between both implementations.

In our experiments, we use a document-oriented decoder: the Docent decoder (Hardmeier et al., 2013; Hardmeier, Nivre, and Tiedemann, 2012). In a nutshell, this decoder moves from a sentence search space to a document search space. It maximizes and computes the translation score for a document as a whole and not only for a sentence. However, Docent also has features that can work at phrase level. In fact, the first step in the document-search of this decoder is equivalent to the SMT system that we described previously.

2.3.1 Semantic Models

The Docent framework also allows to use distributed models as semantic space language models. We want to take advantage of this characteristic and introduce more semantic information in our system by using embeddings trained with the word2vec package (Mikolov et al., 2013a; Mikolov et al., 2013b). Since our goal is to use the embeddings for translation, we train bilingual models following the same strategy as in (Martínez-Garcia et al., 2014): the units used to train the vector models are bilingual pairs of *targetWord_sourceWord*. This kind of vec-

[8] Interpolation weights were trained with the *interpolate-lm.perl* script from Moses and the interpolated language models were binarized afterwards.

tors are useful to capture the information related, not only to the target side or source side words, but also to the translations themselves. For this system, we use the best configuration obtained in (Martínez-Garcia et al., 2014), that is, we train a CBOW architecture using a context window of 5 tokens to get 600-dimensional vectors. The aligned parallel corpus needed to train the models was obtained from the Opus collection and is built up with the OpenSubtitles 2012, 2013, and the Tatoeba and EUbookshop parallel corpora. The final semantic models contain 1,527,004 Catalan_Spanish units and 1,391,022 Spanish_Catalan units. When translating a document, Docent uses these semantic models to estimate an additional score for every phrase that is proportional to the distance among the vectors of that phrase and its local context[9].

3 Evaluation

In the previous section we have introduced three different translation systems: a standard sentence-level SMT system (SMT), a document-level SMT system (DSMT) and a document-level SMT system enriched with additional semantic information (semDSMT). For the shared task we only submitted results with the SMT and semDSMT systems (SMTsub and semDSMTsub systems). However, some problems with the input tokenization were found after the submission. [10] In this section, we report both the results before and after solving this issue. We also found a problem in the integration of the semantic vector models inside the document-oriented decoder (semDSMT systems) that invalidates the results of this system submitted to the task.

Automatic evaluation results for our systems are shown in Table 1. We obtained these results using the Asiya toolkit (Giménez and Màrquez, 2010) for several lexical metrics: WER, PER, TER, BLEU, NIST, GTM2[11], MTRex[12], RGS*[13], Ol[14](Nießen et al., 2000; Tillmann et al., 1997; Snover et al., 2006; Snover et al., 2009; Papineni et al., 2002; Doddington, 2002; Melamed, Green, and Turian, 2003; Denkowski and Lavie, 2012; Lavie and Agarwal, 2007; Lin and Och, 2004) and a normalized arithmetic mean of the lexical metric scores (ULC)(Giménez and Màrquez, 2008). Comparing the SMT and SMTsub systems rows in Table 1 when translating from Catalan to Spanish, it is clear that fixing the tokenization problem in the Catalan test set significantly improves the scores in all the metrics.

Note that, when translating into Spanish, the SMT system outperforms the rest whereas when translating into Catalan the DSMT system is the one with best scores in most metrics. We observe that the differences between the scores of the SMT and DSMT systems are not statistically significant when translating from Spanish to Catalan, but the differences between the scores in the other translation direction are indeed statistically significant, both measured at 95% of confidence level[15]. For example, the BLEU score obtained by the SMT system is 1.32 points higher than DSMT when translating into Spanish, but DSMT has 0.12 points of BLEU more than SMT in the other direction. The similarity between the results for the SMT has two main reasons. On the one hand, the DSMT system departs from the SMT one, so, for an already good translation, such as the ones obtained for tweets, only few changes are applied. On the other hand, the automatic evaluation metrics are not sensitive to the changes due to the context information. It is also important to notice that there exists only one reference. This fact makes more difficult to obtain an accurate evaluation of the translations since correct variations, using synonyms for example, will be scored as wrong translations.

For instance, in the first example in Table 2, we observe how the DSMT obtains a

[9]The local context of a phrase consists of its previous 30 tokens.

[10]There were errors when tokenizing the article form *l'* as well as other elided forms like *'n, d'* or *s'*. Also, we fixed the tokenization of the pronouns that appear after a verb with a dash like in *animar–los* or *donar–nos*.

[11]We use the GTM version with the parameter associated to long matches $e = 2$.

[12]We use the METEOR version using only exact maching.

[13]We use the ROUGE variant which skips bigrams without max-gap-length

[14]Lexical overlap inspired on the Jaccard coeficient for sets similarity.

[15]Significance of the difference between the systems measured for the NIST and BLEU metrics using the implementation of paired bootstrap resampling included in the Moses decoder.

Catalan to Spanish — 140 chars/tweet

System	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	Ol	ULC
SMTsub	20.17	16.40	19.42	68.20	11.22	62.71	78.46	77.31	74.72	65.04
semDSMTsub	25.10	17.09	22.25	63.12	10.93	57.92	76.44	75.56	73.76	58.62
SMT	**14.96**	**11.82**	**14.16**	**76.67**	**12.07**	**71.48**	**84.08**	**81.34**	**82.11**	**78.62**
DSMT	15.74	12.23	14.94	75.35	11.92	69.79	83.38	80.74	81.40	76.75

Catalan to Spanish — free

System	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	Ol	ULC
SMTsub	20.13	16.31	19.38	68.25	11.22	62.71	78.52	77.35	74.76	65.05
semDSMTsub	25.07	17.01	22.21	63.17	10.94	57.92	76.50	75.60	73.80	58.62
SMT	**14.92**	**11.74**	**14.13**	**76.73**	**12.08**	**71.49**	**84.14**	**81.39**	**82.15**	**78.65**
DSMT	15.70	12.15	14.90	75.41	11.92	69.79	83.45	80.79	81.44	76.79

Spanish to Catalan — 140 chars/tweet

System	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	Ol	ULC
SMTsub	14.35	11.25	13.63	77.93	12.04	72.69	53.98	82.19	83.18	66.51
SMT	14.32	11.30	13.58	78.07	12.04	**73.02**	54.08	82.21	83.29	66.62
DSMT	**14.22**	**11.10**	**13.46**	**78.19**	**12.07**	72.96	**54.14**	**82.45**	**83.46**	**67.10**

Spanish to Catalan — free

System	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	Ol	ULC
SMTsub	14.33	11.24	13.61	77.93	12.04	72.69	53.99	82.19	83.20	66.51
SMT	14.31	11.30	13.56	78.06	12.04	**73.02**	54.09	82.21	83.31	66.62
DSMT	**14.20**	**11.09**	**13.44**	**78.19**	**12.07**	72.97	**54.15**	**82.45**	**83.48**	**67.11**

Table 1: Evaluation with a set of lexical metrics for our systems on the Catalan–Spanish language pair. Results include the scores obtained with the raw translations (free) and with the restriction of only considering the first 140 characters per tweet (140 chars/tweet).

better translation than the SMT system with respect to the reference, but actually, both systems obtain a correct translation. There are other examples where the DSMT has a correct translation but it does not match the reference, as shown in the Example 2 from Table 2. In this case, both systems obtain good translations but the SMT translation is closer to the reference since the DSMT uses synonyms for *partido* and *FCB* (*encuentro* and *Barça* respectively). One example where the context information is useful is Example 3 in Table 2 where DSMT uses *cancha* instead of *pista* to translate *pista*, which is a more concrete option since the user account that produced the message is from a famous Spanish basketball team that mostly tweets information about basketball. In the other direction, we found similar phenomena. Example 4 in Table 2 shows again how both systems generate correct translations. In spite of the spelling mistake in the reference (*cumpleix* instead of *compleix*), this time the closest translation to the reference is the one from the SMT system but the DSMT one is still correct.

Most of the problems that we found in our experiments are related to the lack of normalisation of the source and to the decision of keeping the hashtags untranslated. We found several examples where the original tweet is not well written and this produces errors in the translations. For instance, *"Gràcies x ls mencions sobre l'expo #PostPicasso"* where our systems are not able to translate correctly the informal abbreviation *ls*. Regarding the hashtags, we found *"#elmésllegit"* that appears translated as *"#lomásleído"* in the reference but in our systems we decided to preserve the original hashtags.

Example 1: Catalan to Spanish		
Source	Els agents rurals capturen un **voltor comú** a l'Hospitalet	
Reference	Los agentes rurales capturan un ***buitre leonado*** en L'Hospitalet	
SMT	Los agentes rurales capturan un **buitre común** en L'Hospitalet	
DSMT	Los agentes rurales capturan *a* un **buitre leonado** en L'Hospitalet	
Example 2: Catalan to Spanish		
Source	Final del **partit** al Vicente Calderón! ATM 0-0 **FCB**	
Reference	Final del ***partido*** en el Vicente Calderón! ATM 0-0 ***FCB***	
SMT	Final del **partido** en el Vicente Calderón! ATM 0-0 **FCB**	
DSMT	Final del **encuentro** en el Vicente Calderón! ATM 0-0 **Barça**	
Example 3 : Catalan to Spanish		
Source	Aquesta nit, a les 20:30 hores, el IDIDID B visita la **pista** del IDIDID.	
Reference	Esta noche, a las 20:30 horas, el IDIDID B visita la ***cancha*** del IDIDID.	
SMT	Esta noche, a las 20: 30 horas, el IDIDID B visita la **pista** del IDIDID.	
DSMT	Esta noche, a las 20: 30 horas, el IDIDID B visita la **cancha** del IDIDID.	
Example 4 : Spanish to Catalan		
Source	Kim Basinger **cumple hoy** 60 años	
Reference	Kim Basinger ***cumpleix avui*** 60 anys	
SMT	Kim Basinger **compleix avui** 60 anys	
DSMT	Kim Basinger **avui fa** 60 anys	

Table 2: Translation examples of tweets by our different systems in both translation directions: Spanish to Catalan and Catalan to Spanish.

We can also observe that the restriction of 140 characters does not have an important effect in the performance. This is because, for this test set, our systems usually produce tweet translations with a legal length (99.00% from Catalan to Spanish and 99.70% from Spanish to Catalan), and furthermore, among the tweets exceeding the maximum length, the average number of extra characters is less than 6. Notice that it is hard to measure the real length of the tweets since we do not have access to the original messages, instead we have the tweets with the URLs and IDs replaced by their corresponding labels. For the given language pair, our system mostly respect the original length. This is an expected behaviour since the length factor (Pouliquen, Steinberger, and Ignat, 2003) for the Catalan-Spanish language pair is close to 1.

4 Conclusions

We have described the systems developed for the TweetMT shared task: a standard sentence-level SMT system based on Moses and a document-level SMT system based on Docent. We adapted both systems using language models built with tweets. For the document-level SMT system, we considered as context of a tweet the rest of messages from the same user during the same day.

The automatic evaluation of our systems shows that both systems perform similarly. However, it must be taken into account that lexical metrics are not context sensitive and there is only one reference available. As reported in the literature, we found problems with the correctness of the messages and when addressing the problem of translating hashtags as we shown with some examples found during the manual evaluation.

Hashtag translation and normalization of the input are interesting topics for future work especially for extending the system to translate informal tweets. We also consider to implement a pipeline that only takes into account the previous context to simulate an online scenario and compare it with the actual pipeline. Currently we are enhancing the models with the introduction of semantic information using word vector embeddings. In particular, we are customizing the Docent decoder to introduce them at translation time.

References

[Alegria et al.2014] Alegria, I., N. Aranberri, P. R. Comas, V. Fresno, P. Gamallo, L. Padró, I. San Vicente, J. Turmo, and A. Zubiaga. 2014. TweetNorm_es corpus: an annotated corpus for spanish microtext normalization. In *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2274–2278.

[Denkowski and Lavie2012] Denkowski, M. and A. Lavie. 2012. METEOR-NEXT and the METEOR paraphrase tables: Improved evaluation support for five target languages. In *Proc. of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342.

[Doddington2002] Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of the 2nd International Conference on Human Language Technology Research*, pages 138–145.

[Giménez and Màrquez2008] Giménez, J. and L. Màrquez. 2008. A smorgasbord of features for automatic MT evaluation. In *Proc. of the Third Workshop on Statistical Machine Translation*, pages 195–198. ACL.

[Giménez and Màrquez2010] Giménez, J. and L. Màrquez. 2010. Asiya: An open toolkit for automatic machine translation (meta-) evaluation. In *Prague Bulletin of Mathematical Linguistics, 94*, pages 77–86.

[Gotti, Langlais, and Farzindar2013] Gotti, F., P. Langlais, and A. Farzindar. 2013. Translating government agencies' tweet feeds: Specificities, problems and (a few) solutions. In *Proc. of the NACCL 2013*, pages 80–89.

[Hardmeier, Nivre, and Tiedemann2012] Hardmeier, C., J. Nivre, and J. Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proc. of the Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning*, pages 1179–1190.

[Hardmeier et al.2013] Hardmeier, C., S. Stymne, J. Tiedemann, and J. Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proc. of the 51st ACL Conference*, pages 193–198.

[Jehl, Hieber, and Riezler2012] Jehl, L., F. Hieber, and S. Riezler. 2012. Twitter translation using translation-based cross-lingual retrieval. In *Proc. of the 7th Workshop on Statistical Machine Translation. ACL 2012*, pages 410–421.

[Koehn et al.2007] 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 *Proc. of the 45th ACL Conference*, pages 177–180.

[Lavie and Agarwal2007] Lavie, A. and A. Agarwal. 2007. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 228–231.

[Lin and Och2004] Lin, C.-Y and F.J. Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statics. In *Proc. of the 42nd Annual Meeting of the ACL*, pages 605–612.

[Martínez-Garcia et al.2014] Martínez-Garcia, E., C. España-Bonet, J. Tiedemann, and L. Màrquez. 2014. Word's vector representations meet machine translation. In *Proc. of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 132–134.

[Melamed, Green, and Turian2003] Melamed, I.D., R. Green, and J.P. Turian. 2003. Precision and recall of machine translation. In *Proc. of the Joint Conference on HLT-NAACL.*

[Mikolov et al.2013a] Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. In *Proc. of Workshop at ICLR.* http://code.google.com/p/word2vec.

[Mikolov et al.2013b] Mikolov, T., I. Sutskever, G. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS*, pages 3111–3119.

[Munro2010] Munro, R. 2010. Crowdsourced translation for emergency response in haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Crowdsourcing for Translation*, pages 1–4.

[Nießen et al.2000] Nießen, S., F. Och, G. Leusch, and H. Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proc. of the 2nd International LREC Conference*, pages 339–342.

[Och2003] Och, F. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the ACL Conference*, pages 160–167.

[Och and Ney2003] Och, F. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, pages 19–51.

[Papineni et al.2002] Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

[Pouliquen, Steinberger, and Ignat2003] Pouliquen, B., R. Steinberger, and C. Ignat. 2003. Automatic identification of document translations in large multilingual document collections. In *Proc. of the International Conference on Recent Advances in NLP (RANLP-2003)*, pages 401–408.

[Snover et al.2006] Snover, M., B.J. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of the 7th Conference of the AMTA*, pages 223–231.

[Snover et al.2009] Snover, M., N. Madnani, B.J. Dorr, and R. Schwartz. 2009. Fluency, adequacy or HTER? Exploring different human judgments with a tunable MT metric. In *Proc. of the Fourth Workshop on Statistical Machine Translation*, pages 259–268.

[Specia et al.2013] Specia, L., K. Shah, J. G. C. De Souza, and T. Cohn. 2013. QuEst - A translation quality estimation framework. In *Proc. of ACL Demo Session*, pages 79–84.

[Stolcke2002] Stolcke, A. 2002. SRILM – An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, pages 257–286.

[Tiedemann2009] Tiedemann, J. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing (vol V)*, pages 237–248.

[Tiedemann2012] Tiedemann, J. 2012. Parallel data, tools and interfaces in opus. In *Proc. of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, pages 2214–2218. http://opus.lingfil.uu.se.

[Tillmann et al.1997] Tillmann, C., S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP based Search for Statistical Translation. In *Proc. of European Conference on Speech Communication and Technology.*