EHU at TweetMT: Adapting MT Engines for Formal Tweets

EHU en TweetMT: Adaptación de sistemas MT a tuits formales

Iñaki Alegria, Mikel Artetxe, Gorka Labaka, Kepa Sarasola
University of the Basque Country
inaki.alegria@ehu.eus

Resumen: En este trabajo se describe la participación del grupo IXA de la UPV/EHU en la tarea sobre Traducción de Tweets en el congreso de la SEPLN (TweetMT 2015). Se han adaptado dos sistemas previamente desarrollados para la traducción es-eu y eu-es, obteniéndose buenos resultados (mejores que otros publicados previamente). Se describe la recopilación de recursos, la adaptación de los sistemas y los resultados obtenidos.
Palabras clave: traducción automática, SMT, RBMT, tuits, social media

Abstract: This paper describes the participation of the IXA group from the UPV/EHU (University of the Basque Country) in the TweetMT shared task at the SEPLN-2015 conference. We have adapted existing MT engines for the es-eu and eu-es pairs, obtaining good results (better than other experiments reported in previous work). Three main aspects are described: resource compilation, engine adaptation and results.
Keywords: machine translation, SMT, RBMT, tweets, social media

1 Introduction

As the organizers of the workshop say in the home page[1] "*the machine translation of tweets is a complex task that greatly depends on the type of data we work with. The translation process of tweets is very different from that of correct texts posted for instance through a content manager. The texts also vary in terms of structure, where the latter include tweet-specific features such as hashtags, user mentions, and retweets, among others.*" The translation of tweets can be tackled as a direct translation (tweet-to-tweet) or as an indirect translation (tweet normalization to standard text, text translation and, if needed, tweet generation) (Kaufmann and Kalita, 2010).

When analyzing the released development corpus we observed that most of the messages were formal tweets, and we therefore decided to face the problem following the direct approach, adapting previous engines to the structure of these texts. We have adapted three systems:

- an RBMT system named Matxin for the es-eu pair (Mayor et al., 2011). It is well known that automatic measures run on a single reference tend to penalise RBMT systems (rule-based machine translation) compared to SMT systems (statistical machine translation), but we wanted to test the results.
- two state-of-the-art SMT systems, one for the es-eu pair and the other one for the eu-es pair (Labaka, 2010).

2 Resource Compilation from Microtexts

Based on the development set provided, preliminary work was carried out to obtain useful resources to adapt the systems:

- an out-of-vocabulary (OOV) dictionary was obtained for Basque using our Basque morphological analyzer. We observed that the percentage of OOVs was low and only few of them were common in the development set. Even if there are a very small number of entries, we built a bilingual dictionary with the most frequent OOVs (5 entries).
- a dictionary of bilingual hashtags was obtained by aligning hashtags (using a simple program and manual revision) from parallel tweets. After a manual

[1] http://komunitatea.elhuyar.org/tweetmt

review of the pairs with more than two occurrences, a dictionary of 60 pairs was generated.

Tweets of monolingual corpora from previous shared tasks were compiled in order to enrich the language models.

- Corpora from the TweetNorm shared task (Alegria et al., 2014). Initial collection of 227,855 Spanish tweets.
- Corpora from the TweetLId shared task (Zubiaga et al., 2014). 8,562 Spanish tweets and 380 Basque tweets.

In order to increase the low volume of data for the Basque LM we use a corpus of tweets supplied by the CodeSyntax company[2] which has a tweet-oriented service for Basque called UMAP[3]. They identify Twitter accounts that use Basque and compile the tweets from these users. Most of these accounts are multilingual, so language identification was a key next step. We used our language identifier (*LangId*, a free language identifier based on word and trigram frequencies developed by the IXA group of the University of the Basque Country, and which is specialized in recognizing Basque and its surrounding languages (Spanish, French and English)) and filter candidates with high percentage of OOVs (thus, priorizing formal tweets and adding precision to the results obtained from Langid) and compiled a corpus of 454,790 tweets.

3 Adaptation and Tuning of the MT Engines

The RBMT system was adapted to the task manually and two new models for the SMT systems were trained and tuned.

3.1 RBMT

As mentioned, the Matxin system was used for es-eu translation. Matxin is an open source Spanish-to-Basque RBMT engine which follows the traditional transfer model. It consists of three main components: 1) analysis of the source sentence into a dependency tree structure; 2) transfer from the source language dependency tree to a target language dependency structure; and 3) generation of the output translation from the target dependency structure.

Matxin was adapted to the idiosyncratic features of tweets (URLs, hashtags...). For this purpose, the de-formatter module in Matxin (Mayor et al., 2011) was enriched adding the following functions (the deformatter module separates the format information (RTF, HTML, etc.) from the text to be translated, and the plain text is sent to the analysis phase):

- URLs are managed as sentence boundaries
- Hashtags at the begining or at the end of the tweet remain untranslated
- Hashtags inside the text are given for translation. Some will be translated (#Escocia / #Eskozia) while others will remain untranslated (#Hackathon, #Olasdeenergia)
- IDs will receive the same treatment as other named entities

3.2 Corpora for SMT

The adaptation and tuning of the SMT systems was laborious. First of all, the provided development corpus was divided into 3 subsets: training (2,000 pairs), tuning (1,500) and test (500). Because the alignment of the corpus was automatically done, we manually reviewed the training part and observed that the error rate was high. After discarding non-parallel tweets 1,444 pairs remained in the training corpus.

We used a previously compiled parallel corpus for the translation model. This 7.4 million segment corpus was compiled by the Elhuyar Foundation and the University of the Basque Country. It includes public corpora, private corpora and a corpus built by web-as-corpus paradigm (San Vicente and Manterola, 2012). Also, the mentioned training corpus (1,444 pairs from the development corpus in the shared task) was repeated 100 times for the bilingual model (this is not done for the language model because we use interpolation).

For the language model, previous models for Spanish and Basque were retrained adding the corpora described in the previous section.

Table 1 shows the figures for the corpora used.

[2] `http://www.codesyntax.com`

[3] `http://umap.eu`

		Sentences	Tokens (es)	Tokens (eu)
Bilingual	General	7,463,951	118,497,426	94,142,809
	Tweets	1,444	21,022	18,804
Monolingual (es)	General	28,823,939	866,383,394	-
	Tweets	213,141	3,041,837	-
Monolingual (eu)	General	1,290,501	-	14,894,592
	Tweets	454,800	-	6,063,226

Table 1: Figures from the dataset

System	BLEU-c	BLEU	NIST-c	NIST	TER
RBMT(es-eu)baseline	0.1395	0.1629	4.6073	5.1930	0.8824
RBMT(es-eu)enhanced	0.1891	0.2089	5.4024	5.7755	0.7377
SMT(es-eu)baseline	0.2108	0.2257	6.0361	6.4351	0.8116
SMT(es-eu)enhanced	0.2401	0.2635	6.2920	6.7714	0.6550
SMT(eu-es)baseline	0.2348	0.2591	6.2768	6.7493	0.7876
SMT(eu-es)enhanced	0.2826	0.3109	6.9641	7.4827	0.6153

Table 2: Results on the test corpora

3.3 Tuning

The development of the system was carried out using publicly available state-of-the-art tools: the GIZA++ toolkit, the SRILM toolkit and the Moses decoder. More concretely, we followed the phrase-based approach with standard parameters: a maximum length of 80 tokens per sentence, translation probabilities in both directions with Good Turing discounting, word-based translation probabilities (lexical model, in both directions), a phrase length penalty and the target language model. The weights were adjusted using MERT tuning with n-best list of size 100.

For the idiosyncratic features of the tweets we analyzed the errors when the system was applied in the test extracted from the development corpus and we decided to implement the following pre- and post-processing steps:

- Tokenization: special treatment of hyphens ('-') before declension cases of IDs, hashtags, figures, time...
- Post-processing: simple rules for fixing recurrent surface-errors: double hyphen or colon, special symbols (e.g. '¿' is used in Spanish but not in Basque) and similar issues.

4 Results and Discussion

The systems prepared and tuned using the development corpus were directly used to process the test. So we presented two systems for the eu-es pair (RBMT and SMT) and one system (SMT) for the es-eu pair. For these language pairs only another group presented results (3 systems).

Table 2 shows the results on the test corpus provided. We use the most common measures: BLEU and NIST (Doddington, 2002). Our SMT system was the best for the eu-es pair, and the second (very close to the first) for the es-eu pair. As expected, the RBMT system gets lower figures in the metrics (only one reference is supplied) but it is interesting to compare them with previous results.

We want to underline that the results for the es-eu pair are better than previous results reported in some papers (Labaka et al., 2007; Labaka et al., 2014). More specifically, the BLEU figures for the RBMT system in this task range from 0.1429 (baseline) to 0.2089 (improved system) and from 0.2257 (baseline) to 0.2635 (improved system) for SMT; while in the last reference (Labaka et al., 2014) BLEU figures range from 0.0572 to 0.1172 using RBMT and around 0.145 using SMT.

These results are surprising if we consider tweet texts in general, but note that all tweets used in the shared-task are formal and

#	System	Text
1	source	Arranca la segunda mitad GOAZEN! — 0-0 #athlive
	ref.	Hasi da bigarren zatia, aupa!! — 0-0 #athlive
	SMT	Hasi da bigarren zatia GOAZEN! — 0-0 #athlive
	RBMT	Bigarren erdi GOAZEN ateratzen du! — 0-0 #athlive
2	source	Jaume Matas ingresa en prisión URLURLURL
	ref.	Jaume Matas kartzelan sartu dute URLURLURL
	SMT	Jaume Matas kartzelan sartu dute URLURLURL
	RBMT	Jaume Matas espetxean sartzen da URLURLURL
3	source	Retenciones de hasta 7 kilómetros en la AP-8 en Irun: URLURLURL
	ref.	7 kilometroko auto-ilarak AP-8an, Irungo ordainlekuan: URLURLURL
	SMT	7 kilometroko auto-ilarak AP-8 Irunen: URLURLURL
	RBMT	7km-taraino AP-8 Irunen erretentzioak: URLURLURL
4	source	Qu es un OpenSpace? IDIDID 27 de septiembre. URLURLURL
	ref.	Zer da OpenSpace bat? IDIDID irailaren 27an. URLURLURL
	SMT	Zer da OpenSpace? IDIDID 27. URLURLURL
	RBMT	Zer da Openspace bat? Irailaren 27an IDIDID. URLURLURL
5	source	Markel Olano denuncia que Bildu ha decidido actuar en contra de los intereses de los baserritarras #eajpnv URLURLURL
	ref.	Olanok salatu du Bilduk baserritarren interesen kontra egitea erabaki duela #eajpnv URLURLURL
	SMT	Markel Olano salatu du Bilduk jokatzea erabaki du interesen kontra baserritarren #eajpnv URLURLURL
	RBMT	Markel Olanok salatzen du Bilduk baserritarrasen interesen aurka jardutea erabaki duela #eajpnv URLURLURL
6	source	Idoia Mendia reducirá la ejecutiva y asignará tareas a cada miembro URLURLUR
	ref.	Idoia Mendiak exekutiba murriztuko du, bakoitzari zeregin bat emanez URLURLURL
	SMT	Idoia Mendiak eta lan egingo du kide bakoitzari URLURLURL
	RBMT	Idoia Mendiak exekutiboa gutxituko du eta lanak esleituko dizkio kide bakoitzari URLURLURL

Table 3: Examples RBMT/SMT

that the most of them were designed to be multilingual (and so, perhaps, to be easily translated). Therefore, we could say that the task was easier than usual tasks in MT, at least for this language pair.

The good performance of the RBMT system on the formal tweets was expected, as syntax use to be simple in the short texts from Twitter.

In Table 3 there some examples of the results. In sentences #4, #5 and #6 RBMT gets very food translations but in the previous sentences the translations from the SMT system are more precise. In the near future we want to check if combining both techniques improvements can be lead.

We can draw the following general conclusions:

- These results cannot be extrapolated to the general task of translating tweets. Translating informal tweets will be much harder.
- MT can help community managers who manage multilingual Twitter accounts. A Twitter oriented MT post-editing system could be developed and evaluated.

Acknowledgments

This work has been supported by the Spanish MICINN project *Tacardi* (Grant No. TIN2012-38523-C02-01). CodeSyntax company and Elhuyar Foundation have collaborated with us providing several corpora for the translation and language models. Thanks to Josu Azpillaga

(CodeSyntax) and to Iñaki San Vicente, Igor Leturia, Itziar Cortes and Justyna Pietrzak (Elhuyar) for their assistance. We would also like to thank the anonymous referees for their comments and suggestions.

References

Alegria, Inaki, Nora Aranberri, Pere R Comas, Vıctor Fresno, Pablo Gamallo, Lluis Padró, Inaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2014. Tweetnorm_es corpus: an annotated corpus for spanish microtext normalization. In *Proceedings of LREC*.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Kaufmann, Max and Jugal Kalita. 2010. Syntactic normalization of twitter messages. In *International conference on natural language processing, Kharagpur, India*.

Labaka, Gorka. 2010. Eusmt: incorporating linguistic information into smt for a morphologically rich language. its use in smt-rbmt-ebmt hybridation. *Lengoaia eta Sistema Informatikoak Saila (UPV-EHU). Donostia. 2010ko martxoaren 29a*.

Labaka, Gorka, Cristina España-Bonet, Lluís Màrquez, and Kepa Sarasola. 2014. A hybrid machine translation architecture guided by syntax. *Machine Translation*, 28(2):91–125.

Labaka, Gorka, Nicolas Stroppa, Andy Way, and Kepa Sarasola. 2007. Comparing rule-based and data-driven approaches to spanish-to-basque machine translation.

Mayor, Aingeru, Iñaki Alegria, Arantza Díaz De Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. Matxin, an open-source rule-based machine translation system for basque. *Machine translation*, 25(1):53–82.

San Vicente, Inaki and Iker Manterola. 2012. Paco2: A fully automated tool for gathering parallel corpora from the web. In *LREC*, pages 1–6.

Zubiaga, Arkaitz, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Vıctor Fresno. 2014. Overview of tweetlid: Tweet language identification at sepln 2014. *TweetLID@SEPLN. TweetLId workshop at SEPLN Conference. ceur-ws.org/Vol-1228/*.