Dealing with Sign Language Morphemes in Statistical Machine Translation

Guillem Massó, Toni Badia
Grup de Lingüística Computacional (GLiCom)
Universitat Pompeu Fabra, Roc Boronat 138, Barcelona, Spain
E-mail: guillem.masso@upf.edu, toni.badia@upf.edu

Abstract

The aim of this research is to establish the role of linguistic information in data-scarce statistical machine translation for sign languages using freely available tools. The main challenge in statistical machine translation is the scarcity of suitable data, and this problem becomes more pronounced in sign languages. The available corpora are small, usually not domain-specific, and their annotation conventions can vary considerably. Elaborating our own corpus is a very time-consuming task and the amount of data that we can obtain is even more reduced. Under these conditions, morpho-syntactic information helps to improve statistical machine translation results, but there are not linguistic processing tools for sign languages. We have managed to improve translations from Catalan to Catalan Sign Language by using factored models in an open source translation system with basic linguistic information such as the lemma or an annotation tier tag. Furthermore, this allows us to deal with sign language morphemes in a more systematic way.

1. Introduction

Nowadays, there is an increasing interest in data-driven approaches in machine translation (DDMT), either statistical (SMT) or example-based (EBMT): their development is less time-consuming and they are more scalable than rule-based approaches, although a considerable amount of data is required to create bilingual corpora. So, the main challenge is to set up a suitable parallel corpus large enough. Problems in SMT due to scarce resources, which are endemic in sign languages (SLs), have also been detected in oral languages (OLs). One of the suggested solutions to improve translation results is to use morpho-syntactic information (Nießen and Ney, 2004). This is a good solution if there are linguistic processing tools for the analysed languages: once more, SLs are at a disadvantage. However, these tools can be used for the OL corpus part, and other alternatives must be found for the SL analysis. In this work, we propose two solutions: the use of plain glosses as lemmas of inflected forms, and the use of annotation tier names as tags in a more syntactical approach. This linguistic information is integrated at the word level using factored models of Moses, an open source SMT system (Koehn et al., 2007).

The remainder of the paper is organized as follows. In section 2, we give a brief overview of related work in DDMT and the use of morpho-syntactic information. In section 3, we present our parallel corpus for the language pair Catalan and Catalan Sign Language (LSC). Section 4 describes the experiments carried out and section 5 discusses the results and evaluation. Finally, section 6 outlines the main conclusions of the work.

2. Related work

As for sign language MT (SLMT), and as far as we are aware, there are four main research groups working on DDMT. Stein, Bungeroth and Ney (2006) use a phrase-based SMT system for the language pair German and German SL (DGS). The SL corpus is annotated with glosses, including all important grammar features. Their research is focused on morpho-syntactic pre and post-processing enhancement. In the pre-processing step, German is analysed by a parser, and part-of-speech (POS) information is used to transform nouns into stem forms, split compound words and delete German POS not used in DGS. In the post-processing step, marked positions of discourse entities are added from a database. Some deleted information about emphasis and comparative degree is added as well. Therefore, morpho-syntactic information is not used during the translation process.

The research of Morrissey and Way (2007) focuses on EBMT. The ATIS corpus (Bungeroth et al., 2008) was translated from English to Irish SL (ISL) to be used as data set. The SL data are annotated with glosses but without non-manual or phonetic feature detail, and no morpho-syntactic information is used.

The two aforementioned groups have collaborated in Stein et al. (2007) and Morrissey et al. (2007) to translate from SL to OL with SL recognition. Although their research does not focus on morpho-syntactic improvements in SLMT, some interesting issues are raised. The main one concerns the handicap of lacking SL parsers, since morpho-syntactic information usually reduces errors. However, the authors consider that adding features such as the hand tracking position in pointing signs is comparable to adding POS information. They also suggest that "other features are likely to improve the error rates as well and should be investigated further" (Stein et al., 2007).

Su and Wu (2009) go beyond and use a treebank, a bilingual dictionary and a translation memory to convert the Chinese syntactic structure with thematic role information into the corresponding structure in Taiwanese SL. Thematic roles also allow them to deal with

agreement verbs by identifying verb arguments and providing movement directions. However, the authors highlight that the proposed system hardly deals with non-manual features, although this issue would be the next step in their research.

San-Segundo et al. (2008) work on speech recognition and MT from Spanish to Spanish SL (LSE). They compare a rule-based MT system with a SMT system. The rule-based system obtains the best results: on one hand, the restricted domain (a service for renewing identity cards) makes it possible to develop a complete set of rules with reasonable efforts, and on the other hand, the statistical system cannot be trained properly due to the reduced amount of data. They also collaborated with the Aachen group (D'Haro et al., 2008) to improve the sign language model using information retrieval from the Web.

3. Corpus

Nowadays, there is not any available corpus in LSC which could be used for MT, so we have created a small corpus on the weather report domain. This is a restricted domain with a limited vocabulary that allows us to obtain reasonable results with scarce resources. The original Catalan texts were retrieved from the Catalan Weather Service website (Servei Meteorològic de Catalunya[1]) and translated by a native Deaf signer. Catalan sentences were analysed with the freely accessible tagger CatCG[2] (Alsina et al., 2002) to obtain lemmas and POS, and they were manually revised. The recorded LSC sentences were annotated with iLex (Hanke & Storz, 2008), which allows a greater control over the annotation process thanks to its lexical database.

We were especially interested in morphemes containing adverbial and aspectual information. In order to systematically annotate these linguistic features, the gloss tier contains plain glosses, which we will consider lemmas, and there are separated tiers for mouth morphemes and for movement morphemes. Regarding annotation, the currently available guidelines (Neidle, 2002; Nonhebel, Crasborn & van der Kooij, 2004) do not offer a suitable description for the analysed LSC morphemes, so specific tags have been created. However, the important thing is not the tag assigned, but the fact that morphemes are individualised and classified.

We made two sets from the annotation files. Both sets have added factors with linguistic information, but they differ in SL morphemes representation. In set *1*, morphemes are attached to glosses and the lemma is a factor, represented by the plain gloss. In set *2*, morphemes are independent tokens and the added factor is the annotation tier name. It can be seen in the next example, where the vertical bar separates factors, *ct* stands for the mouth morpheme *cheeks puffed and tense*, and *f* stand for the movement morpheme *fast movement*. This example means 'heavy rain':

Set *1*: RAIN:ct:f|RAIN
Set *2*: RAIN|gloss ct|mouth f|movement

Statistics of the bilingual corpus are shown in Table 1. Notice that there are not lemmas in set *2* because there is not form variation.

		Catalan	LSC (Set 1)	LSC (Set 2)
Training	Sentences	153		
	Running words	1967	1520	1930
	Vocabulary	282	220	182
	Lemmas	241	162	n/a
	Singleton words	87	77	50
	Singleton lemmas	66	46	n/a
Test	Sentences	46		
	Running words	449	376	479
	Vocabulary	164	130	116
	Lemmas	146	102	n/a
	Singleton words	88	64	45
	Singleton lemmas	70	41	n/a
	OOV words	10	5	2
	OOV lemmas	7	2	n/a

Table 1: Statistics of the bilingual corpus with two annotation sets for Catalan Sign Language (LSC).

4. Experiments

The system used is Moses (Koehn et al., 2007), an open source toolkit for SMT. Moses relies on SRILM (Stolcke, 2002) to create language models (LM) of the target language, and on GIZA++ (Och & Ney, 2003) for the alignment process. This system enables the integration of additional information at the word level using factored models. As for the OL, we use the lemma and the POS as added factors. As for the SL, the added factor is the lemma in set *1*, and the annotation tier in set *2*, as mentioned in the previous section.

In previous tests, we noticed that using a smaller training set plus a development set to tune the translation models gives worse results than using a bigger training set without tuning, probably due to the small amount of data. In the end we decided to train and tune the system with the whole training set in order to optimize the results. The LM was also improved by considering all the available sentences of the training and test sets. It is important to highlight that the system creates one LM for each factor of the target language. The built LMs are based on tri-grams.

Given that the aim of these experiments is to evaluate the role of linguistic information, the factors of source and target languages are combined in different ways. As for the source language, translations are from: form, lemma, form + lemma, lemma + POS, form + lemma + POS. As for the target language, translations are to: form, form +

[1] http://www.meteo.cat/mediamb_xemec/servmet/index.html
[2] http://www.glicom.upf.edu/projectes/catcg

factor (lemma or annotation tier). Altogether, there are ten translations per set.

5. Results

5.1 Machine evaluation

All the translations have been evaluated with the NIST and BLEU metrics, as can be seen in Table 2. The most relevant fact is that translations with an added factor for the target language (set *b*) are considerably better than translations to only the target form (set *a*). As for the source language factors, it is not always clear that they can improve the translation. Differences between set *1* and set *2* depend on factors as well, and the two metrics are not always coherent.

		Set *1*		Set *2*	
		NIST	**BLEU**	**NIST**	**BLEU**
Set *a*	F→F	5.0682	0.4427	5.2071	0.3967
	L→F	5.6373	0.4958	5.3476	0.4307
	F+L→F	5.5198	0.4700	5.2515	0.3908
	L+POS→F	**5.7826**	**0.5059**	5.2255	0.3939
	F+L+POS→F	5.4497	0.4596	**5.4658**	**0.4378**
Set *b*	F→F+AF	6.8178	0.6294	6.3251	0.6951
	L→F+AF	6.6842	**0.6373**	6.3809	**0.7245**
	F+L→F+AF	6.8968	0.6271	6.1389	0.6783
	L+POS→F+AF	6.5717	0.6172	**6.4224**	0.7111
	F+L+POS→F+AF	**6.9004**	0.6234	6.4110	0.7143

Table 2: Machine evaluation results.
(F = form, L = lemma, AF = added factor)

In subset *1a*, the worst results are for translations from the surface form, and the maximum improvement is of 0.7144 in NIST and 0.0632 in BLEU by using the lemma and the POS. The second best score is for translations from the lemma. Nevertheless, if the three factors are used (form + lemma + POS), the second worst result is obtained. On the other hand, in subset *2a*, the latter combination is the best, and the second best score is again for the lemma. The differences among the other three options are rather low. The score variability in subset *2a* is of 0.2587 in NIST and 0.0470 in BLEU. In general, scores are better in subset *1a* than in subset *2a*.

In subset *1b*, the worst results are for translations from the lemma and the POS. In the other cases, the metrics are not coherent. In NIST, the best scores are for (in this order): form + lemma + POS, form + lemma, form, lemma. In BLEU, the order is inverted. The score variability is of 0.3287 in NIST and 0.0201 in BLEU. In subset *2b*, translations from form + lemma obtain the worst results, followed by translations from only the form. The best scores in NIST are for lemma + POS, form + lemma + POS and lemma. In BLEU, for lemma, form + lemma + POS and lemma + POS. The score variability is of 0.2835 in NIST and 0.0462 in BLEU. Within the set *b*, NIST scores are higher in subset *1b*, while BLEU scores are higher in subset *2b*.

While the maximal improvement by combining source factors has been of 0.7144 in NIST and 0.0632 in BLEU, the improvement by adding one target factor has been of 0.7891-1.7496 in NIST and 0.1113-0.3172 in BLEU. This is probably due to the fact that the system has two related LMs, which improves the quality of the target sentences, although the LM had already been optimized. Considering these results, the improvement of the LM seems to be more important than the improvement of the translation model. In addition, it is difficult to find clear patterns for the role of source factors in the translation process.

5.2 Human evaluation

Unfortunately, it was not possible to conduct a human evaluation by native Deaf signers. Nevertheless, it is interesting to analyse some translation results in order to clarify the role of source factors and the differences between set *1* and set *2*. Four translations were chosen: form → form, form + lemma + POS → form, form → form + factor, form + lemma + POS → form + factor. We evaluated the sentences from 1 (wrong) to 5 (correct) and we noticed than 27 sentences had been correctly translated in all the cases. These sentences fulfil two conditions: they have been seen in the training set and their length is equal or lower than 10 words. As their translation difficulty is low, we will analyse the other 19 sentences, 3 of which are seen sentences longer than 10 words and 16 are not seen sentences. The number of sentences for each score and the average per sentence are shown in Table 3.

		Score					
		5	**4**	**3**	**2**	**1**	**Average**
Set *1*	F → F	3	4	6	5	1	3.11
	F+L+POS → F	5	1	8	5	0	3.32
	F → F+AF	4	2	8	5	0	3.26
	F+L+POS → F+AF	5	4	6	4	0	**3.53**
Set *2*	F → F	3	0	4	9	3	2.53
	F+L+POS → F	1	5	5	7	1	2.89
	F → F+AF	1	1	6	9	2	2.47
	F+L+POS → F+AF	3	4	7	4	1	**3.21**

Table 3: Human evaluation results by number of sentences for each score.
(F = form, L = lemma, AF = added factor)

Concerning the differences between the two sets, the scores of set *1* are clearly higher than the corresponding ones of set *2*. We have noticed that set *2* has more syntactic errors due to incorrect positions assigned to morphemes and to wrong gloss-morpheme combinations. As for the factors considered, the best results are obtained with all of the factors of both languages. It is important to highlight that the improvement by adding the source

factors is higher than that by adding the target factor, contrary to what machine evaluation shows.

6. Conclusions

Although a complete human evaluation by native Deaf signers would be necessary, we can assert that factored models with linguistic information for both source and target languages improve the results of statistical SLMT. Regarding the SL, complex morpho-syntactic analyses are not indispensable, but simple information from annotation files can be used in an efficient way. Furthermore, this allows us to deal with SL morphemes, which are usually ignored in SLMT. The analysis of the results shows that the best solution of the two proposals is to attach morphemes to glosses and to use plain glosses as lemmas, which are used as added factors. The other solution, considering morphemes as independent tokens, can generate additional syntactic errors.

7. Acknowledgements

This research was supported by the Commission for Universities and Research of the Department of Innovation, Universities and Enterprise of the Catalan Government (Generalitat de Catalunya) and by the European Social Fund.

8. References

Alsina, À., Badia, T., Boleda, G., Bott, S., Gil, À., Quixal, M., Valentín, O. (2002). CATCG: A General Purpose Parsing Tool Applied. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, Spain, 3, pp. 1130—1134.

Bungeroth, J., Stein, D., Dreuw, P., Ney, H., Morrissey, S., Way, A., van Zijl, L. (2008). The ATIS Sign Language Corpus. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, pp. 2943—2946.

D'Haro, L.F., San-Segundo, R., Córdoba, R., Bungeroth, J., Stein, D., Ney, H., (2008). Language Model Adaptation for a Speech to Sign Language Translation System using Web Frequencies and a MAP Framework. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*, Brisbane, Australia, pp. 2199—2202.

Hanke, T., Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora (LREC 2008)*, Marrakech, Morocco, pp. 64—66.

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., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic, pp. 177—180.

Morrissey, S., Way, A. (2007). Joining Hands: Developing a Sign Language Machine Translation System with and for the Deaf Community. In *Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision & Hearing Impairments: Assistive Technology for all Ages (CVHI 2007)*, Granada, Spain.

Morrissey, S., Way, A., Stein, D., Bungeroth, J., Ney, H. (2007). Combining Data-Driven MT Systems for Improved Sign Language Translation. In *Proceedings of the Machine Translation Summit XI (MT'07)*, Copenhagen, Denmark, pp. 329—336.

Neidle, C. (2002). SignStream™ Annotation: Conventions used for the American Sign Language Linguistic Research Project. *American Sign Language Linguistic Research Project Reports*, 11.

Nießen, S., Ney, H. (2004). Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics*, 30(2), pp. 181—204.

Nonhebel, A., Crasborn, O., van der Kooij, E. (2004). *Sign language transcription conventions for the ECHO Project*. Radboud University Nijmegen.

Och, F.J., Ney, H.(2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1), pp. 19—51.

San-Segundo, R., Barra, R., Córdoba, R., D'Haro, L.F., Fernández, F., Ferreiros, J., Lucas, J.M., Macías-Guarasa, J., Montero, J.M., Pardo, J.M. (2008). Speech to sign language translation system for Spanish. *Speech Communication*, 50(11-12), pp. 1009—1020.

Stein, D., Bungeroth, J., Ney, H. (2006). Morpho-Syntax Based Statistical Methods for Automatic Sign Language Translation. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation (EAMT 2006)*, Oslo, Norway, pp. 169—177.

Stein, D., Dreuw, P., Ney, H., Morrissey, S., Way, A. (2007). Hand in Hand: Automatic Sign Language to Speech Translation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, Skövde, Sweden, pp. 214—220.

Stolcke, A. (2002). SRILM – An Extensible Language Model Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, USA, vol. 2, pp. 901—904.

Su, H.-Y., Wu, C.-H. (2009). Improving Structural Statistical Machine Translation for Sign Language with Small Corpus Using Thematic Role Templates as Translation Memory. *IEEE Transactions on Audio, Speech and Language Processing*, 17(7), pp. 1305—1315.