Experimenting the use of *catenae* in Phrase-Based SMT

Manuela Sanguinetti
Dipartimento di Informatica
Università di Torino (Italy)
Corso Svizzera, 185, 10149 Torino
manuela.sanguinetti@unito.it

Abstract

English. Following recent trends on hybridization of machine translation architectures, this paper presents an experiment on the integration of a phrase-based system with syntactically-motivated bilingual pairs, namely the so-called *catenae*, extracted from a dependency-based parallel treebank. The experiment consisted in combining in different ways a phrase-based translation model, as typically conceived in Phrase-Based Statistical Machine Translation, with a small set of bilingual pairs of such catenae. The main goal is to study, though still in a preliminary fashion, how such units can be of any use in improving automatic translation quality.

Italiano. *L'integrazione di conoscenza linguistica all'interno di sistemi di traduzione automatica statistica é un trend diffuso e motivato dal tentativo di combinare le migliori caratteristiche dei sistemi basati su regole con approcci puramente statistici e basati su corpora. Il presente lavoro si inserisce all'interno di queste ricerche e costituisce uno studio preliminare sull'applicazione di una nozione sintattica basata su dipendenze, quella delle cosiddette "catenae", all'interno di una tipica architettura di traduzione statistica.*

1 Introduction

The hybridization of machine translation systems in order to benefit from both statistical-based and linguistically-motivated approaches is becoming a popular trend in translation field. Such trend is well described in a number of surveys (Costa-Jussá and Farrús, 2014; Costa-Jussá and Fonollosa, 2015) and witnessed by recent initiatives in NLP community, such as the HyTra workshop series[1]. The motivations to this choice can be manifold, but essentially lie in the need to either reduce the costs - both in terms of time and resources - of building a fully rule-based system, or to integrate statistical models or SMT outputs with linguistic knowledge, as this could be useful to capture complex translation phenomena that data-driven approaches cannot handle properly.

Such phenomena are often called translational divergences, or even *shifts* (Catford, 1965), and usually involve a large number of linguistic and extra-linguistic factors.

Our main research interest is the study of such shifts, in particular from a syntactic point of view, and of how such linguistic knowledge could be of any use to overcome the current shortcomings in machine translation.

The preliminary experiment presented here is therefore guided by the second motivation mentioned above: our basic assumption is that supplementing translation models in classical Phrase-Based Statistical Machine Translation (PBSMT) with syntactically-motivated units extracted from parallel treebanks can lead to improvements in machine translation accuracy. This was already demonstrated, for example, in Tinsley (2009), where syntactic contituents were used to improve the translation quality of a PBSMT system. However, instead of a constituency paradigm, we focused on a more dependency-oriented syntactic unit, namely the one of *catena*. The choice of a dependency-paradigm in general is mainly dictated by the acknowledged fact that dependencies can better represent linguistic phenomena typical of morphologically rich and free-word order languages (see e.g. (Covington, 1990; Goldberg et al., 2013)). On the other hand, to capture translation shifts of various nature, it is necessary to consider a syntactic unit that goes beyond the single

[1] http://www.hyghtra.eu/workshop.html

node, as also recently pointed out, e.g., in Deng et al. (2015); hence the introduction of the notion of *catena* in our study.
In order to verify our assumption, we carried out a preliminary experiment performing several translation tasks, with Italian and English as language pair. For this purpose, a typical phrase-based SMT system was built, using for training the translation model various combinations of baseline SMT configurations and pairs of catenae automatically extracted from a parallel treebank, i.e. ParTUT, and then automatically aligned.

The remainder of this paper is thus organized as follows: Section 2 introduces the notion of catena, in Section 3 we describe our use of catenae in this experiment, while in Section 4 we describe the training configurations chosen and discuss the results.

2 Catenae: a brief introduction

A large number of contributions, in MT, provided some hints on the need to infer complex translational patterns - often encoded by one-to-many or many-to-many alignments - by including a more extensive hierarchical notion that goes beyond the mere word level. In constituency frameworks, such notion is fully covered by syntactic phrases, or constituents, while in dependency contexts - where this is not explicitly defined - a number of different approaches have been proposed to tackle the problem; Ding and Palmer (2004) (and follow-up works) proposed the extraction and learning of the so-called *treelets*, which refer to any arbitrary dependency subgraph that does not necessarily goes down to some leaf. Recently though, a new unit type has been defined in dependency framework, which, to a certain extent, linguistically justifies and formalizes the abovementioned notion of treelet (originally conceived for computational purposes only). This is the notion of *catena* (Latin for "chain", pl. *catenae*). In Kiss (2015), a catena is defined as:

> *a single w(ord) or a set of w(ords) C such that for all w in C, there is another w' in C that either immediately dominates or is dominated by w. According to this definition, any given tree or any given subtree of a tree qualifies as a catena.*

As a result, catena is claimed to be more inclusive than constituents, as it does not require the unit to include all the nodes that are dominated. Because of the dominance constraint, however, it cannot be compared to a string either.

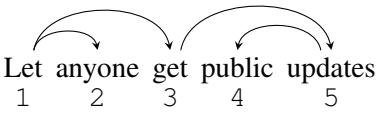

Figure 1: Example of catena.

Figure 1 shows an example of a sentence represented in an unlabelled dependency graph where each word is assigned an identifier (`1, 2, 3, 4, 5`). In the sentence, 15 distinct catenae can be identified (including single nodes)[2]: `[1]`, `[2]`, `[3]`, `[4]`, `[5]`, `[1 2]`, `[1 3]`, `[3 5]`, `[4 5]`, `[1 2 3]`, `[1 3 5]`, `[3 4 5]`, `[1 2 3 5]`, `[1 3 4 5]`, and `[1 2 3 4 5]` (i.e. the whole dependency graph).

A catena may thus include both contiguous and non-contiguous sequences of words, such as *Let get* or *Let get updates*; however, this is not the case for the string "*Let anyone get public*", since there is no direct path to the word "*public*".

The usefulness of catenae in theoretic accounts of complex linguistic phenomena has already been widely shown in literature (Osborne, 2005; Osborne et al., 2011; Osborne and Putnam, 2012; Simov and Osenova, 2015). And to our knowledge, only a few NLP studies (even beyond the bare MT field) exploited this syntactic unit for some practical purpose. The only study we are aware of so far is that of Maxwell et al. (2013), who present an approach based on catenae to *ad hoc* Information Retrieval. It is our opinion, however, that even translation issues can be tackled by integrating such inclusive notion; catenae can be used, for example, to explain and properly identify those cases of one-to-many or many-to-many correspondences, typical of several translation shifts, such as different underlying syntactic structures, MWEs or idioms. For this reason we attempted to exploit them in this experimental study, among other purposes.

3 Catenae extraction and alignment

The first preprocessing step in this experiment consists in the extraction of the possible catenae

[2]In accordance with the convention used in (Osborne et al., 2012), the words that form a catena are listed in a left-to-right order, following their linear order in the sentence.

from parse trees of a parallel treebank. The resource we used for this purpose is ParTUT, a recently developed parallel treebank for Italian, English and French[3] (Sanguinetti and Bosco, 2014). The whole treebank currently comprises an overall amount of 148,000 tokens, with approximately 2,200 sentences in the Italian and English sections respectively, and 1,050 sentences for French.

For this experiment, we used the Europarl section of the treebank, retaining only the sentence pairs that have a direct correspondence (1:1), hence using a set of 376 pairs with an average of 10K tokens per language. To each monolingual file, formatted in CoNLL, of this parallel set we then applied the script for the extraction of catenae.
The script basically performs a depth-first search into the dependency tree, and for each node *w* recursively detects all the possible catenae starting from *w* to the nodes that, directly or indirectly, it dominates. The output file thus provides for each sentence a sequence of such catenae (one per line).

Although the parallel sentences perfectly match with each other, this is not obviously the case for catenae as well. For this reason we carried out a further preprocessing step that entailed the automatic alignment of the output English and Italian files containing such catenae. The alignment was performed considering catenae as if they were sentences, thus using the Microsoft Sentence Bilingual Aligner[4] (Moore, 2002) as alignment tool, and setting a high-probability threshold (0.99) in order to have a more accurate - though far more reduced[5] - pairs of parallel catenae. The set obtained in this step consists of about 1,700 pairs (set A), which was further filtered to obtain a separate subset of pairs - 778 in total - where each catena has a 7-token maximum length (set B). Such subset was created so as to be used in a different training configuration during the translation step (see next section).

Once extraction and alignment steps were completed, we proceeded with the translation tasks, as detailed in the next section.

[3] http://www.di.unito.it/~tutreeb/partut.html

[4] Downloadable here: http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/

[5] The extremely smaller amount of aligned catenae may also be explained by the fact that the order in which the source and target sentences (and catenae, in our case) are listed impacts on the amount and quality of the final alignments.

4 Using catenae in PBSMT

To perform the task, we used Moses (Koehn et al., 2007) as translation toolkit, and set up the system so as to train multiple models, that correspond to the baseline model and to the baseline model augmented with catenae in two different ways.

4.1 Data

Because of its size and availability, the Europarl-v7 parallel corpus (Koehn, 2005) was used for training and testing the system.

To train the baseline translation model, we used a set of 100K parallel sentences, that, however, reduced to an amount of approximately 85K after cleaning up the corpus (we just retained the sentence pairs of up to a 50-tokens length), while we retrieved a far smaller set for tuning (850 sentences) and a set of 1000 sentences for testing.

As we built a system for both translation directions, the language model was computed for both languages using the entire monolingual sets on the English and Italian sides of the corpus (around 1.9M sentences each).

4.2 Experimental setup

The baseline system was built using the basic phrase-based model, which typically does not make any explicit use of linguistic information.
For language modeling, we opted for the trigram option using the IRSTLM toolkit (Federico et al., 2008).
The translation model was computed using the default settings provided by the system guidelines. Word alignment was performed with GIZA++ (Och and Ney, 2003) and 'grow-diag-final-and' as symmetrization heuristic, while a default length of 7 was kept for phrases.
This model, however, was also adapted so as to be configured with three different options:

- to be trained with phrase pairs only (BASELINE)
- to be trained by adding to the baseline training corpus the set A of aligned catenae described in Section 3 (BASELINE+TRAIN)
- to be trained with a combination of multiple sources, i.e. extending Moses' phrase table with the set B of aligned catenae mentioned in Section 3 (BASELINE+CAT)

source sentence	*sia l' Islam che il mondo cristiano sostengono i diritti delle donne*
reference	*for Islam and Christianity both uphold the rights of women*
BASELINE	*both Islam that the Christian world are the rights of women*
BASELINE+TRAIN	*both Islam and Christianity support the rights of women*

Table 1: Translation example.

The second and third configurations were obtained using a simple approach, i.e. concatenating the bilingual catenae *a*) to the training files (BASELINE+TRAIN), and *b*) to the list of the corpus-extracted phrase pairs (BASELINE+CAT).

The final translation outputs were then evaluated with BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) scores, and results are discussed in the next section.

4.3 Results

The findings emerged from the final evaluation, as also reported in Table 2, show very different results both according to the type of configuration used and to the translation direction. However, from such diversified outputs, relevant data can be highlighted.

Such relevance mainly consists in the improvement of translation quality when simply augmenting the training corpus with other external data (BASELINE+TRAIN). As a matter of fact, although such improvement is far from significant in terms of BLEU score in Italian-to-English translation, its NIST counterpart, together with the overall quality of English-to-Italian translation show more encouraging results, with an increase from 6.2410 to 6.2599 in NIST score for the first case, and a 0.02 and 0.03 points in BLEU and NIST scores respectively, for the the second one. Table 1 shows an example translation of an Italian sentence comparing BASELINE and BASELINE+TRAIN outputs.

A small improvement is also reported in the NIST score of the Italian-to-English model when adding a set of bilingual catenae into the phrase table (BASELINE+CAT). This case as well may not be particularly significant in itself, though however encouraging, considering the small amount of data that was added with respect to the baseline system. On the other hand, such set does not seem to affect at all the English-to-Italian model. As a matter of fact, it produces the same hypothesis translation than the one produced with the baseline configuration, and both translations are reported to have a lower translation quality with respect to the first system pair, despite the same amount of training data was used in both directions, even for the language modeling. Such result can be probably explained with some error in the tuning process, while the overall lower quality may be explained, we hypothesize, as an effect of translating into a morphologically richer language - though more in-depth studies should be carried out to support this hypothesis.

		BLEU	**NIST**
It-En	BASELINE	0.2610	6.2410
	BASELINE+TRAIN	**0.2621**	**6.2599**
	BASELINE+CAT	0.2609	6.2582
En-It	BASELINE	0.2241	5.9161
	BASELINE+TRAIN	0.2427	6.2194
	BASELINE+CAT	0.2241	5.9161

Table 2: Experimental evaluation of Italian-to-English and English-to-Italian translation quality under a baseline PBSMT system, and other two PBSMT systems integrated with catenae.

5 Conclusions

The paper presented a small experiment on the combined use of linguistic knowledge - in the form of syntactically-motivated translation units - and statistical model provided by state-of-the-art machine translation techniques. The results reported here are to be considered preliminary, as they suffer from the absence of systematic procedures and data that could not have been applied so far due to lack of time and proper resources. Still, considering these shortcomings, translation evaluation, at least in one direction, produced promising results. There is however a lot of work to do to under this respect in order to effectively improve translation quality with the help of such linguistic information; for example by scaling up this experiment using a larger set of external data, or using different training configurations, so as to have multiple

sources of comparison for final assessments and considerations.

References

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), May.

John C. Catford. 1965. *A Linguistic Theory of Translation: An Essay on Applied Linguistics*. Oxford University Press.

Marta R. Costa-Jussá and Mireia Farrús. 2014. Statistical machine translation enhancements through linguistic levels: a survey. *ACM Computing Surveys (CSUR)*, 46:1–28, January.

Marta R. Costa-Jussá and José A.R. Fonollosa. 2015. Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*, 32:3—10, July.

Michael A. Covington. 1990. Parsing discontinuous constituents in dependency grammar. *Comput. Linguist.*, 16(4):234–236, December.

Dun Deng, Nianwen Xue, and Shiman Guo. 2015. Harmonizing word alignments and syntactic structures for extracting phrasal translation equivalents. In *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST9)*, pages 1–9.

Duan Ding and Martha Palmer. 2004. Automatic learning of parallel dependency treelet pairs. In *Proceedings of the 1rst International Joint Conference on Natural Language Processing (IJCNLP-04)*, pages 233–243.

George Doddington. 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*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. Irstlm: an open source toolkit for handling large scale language models. In *9th Annual Conference of the International Speech Communication Association (INTERSPEECH '08)*, pages 22–26.

Yoav Goldberg, Yuval Marton, Ines Rehbein, and Yannick Versley, editors. 2013. *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*. Association for Computational Linguistics, Seattle, Washington, USA, October.

Tibor Kiss. 2015. *Syntax - Theory and Analysis. Vol. 2*. Walter de Gruyter GmbH & Co KG, Berlin.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*, pages 177–180.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*.

K. Tamsin Maxwell, Jon Oberlander, and W. Bruce Croft. 2013. Feature-based selection of dependency paths in ad hoc information retrieval. In *ACL '13*, pages 507–516.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA-02)*, pages 135–144.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 1(29):19–51.

Timothy Osborne and Michael Putnam. 2012. Constructions are catenae: Construction grammar meets dependency grammar. *Cognitive Linguistics*, 23:165–215.

Timothy Osborne, Michael Putnam, and Thomas Gross. 2011. Bare phrase structure, label-less trees, and specifier-less syntax: Is minimalism becoming a dependency grammar? *The Linguistic Review*, 28:315–364.

Timothy Osborne, Michael Putnam, and Thomas Gross. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15:354–396.

Timothy Osborne. 2005. Beyond the constituent: A dependency grammar analysis of chains. *Folia Linguistica*, 39:251–297.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manuela Sanguinetti and Cristina Bosco. 2014. Parttut: The turin university parallel treebank. In Roberto Basili, Cristina Bosco, Rodolfo Delmonte, Alessandro Moschitti, and Maria Simi, editors, *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, pages 51–69.

Kiril Simov and Petya Osenova. 2015. Catena operations for unified dependency analysis. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 320–329.

John Tinsley. 2009. Resourcing machine translation with parallel treebanks. phd thesis.