A Richly Annotated, Multilingual Parallel Corpus for Hybrid Machine Translation

Eleftherios Avramidis[4], Marta R. Costa-jussà[1], Christian Federmann[4], Maite Melero[1], Pavel Pecina[2], Josef van Genabith[3]

[1] Barcelona Media, Barcelona, Spain
[2] Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic
[3] DCU Dublin City University, Dublin, Ireland
[4] DFKI GmbH, Berlin & Saarbrücken, Germany
eleftherios.avramidis@dfki.de, marta.ruiz@barcelonamedia.org, cfedermann@dfki.de,
maite.melero@barcelonamedia.org, pecina@ufal.mff.cuni.cz, josef@computing.dcu.ie

Abstract

In recent years, machine translation (MT) research has focused on investigating how hybrid machine translation as well as system combination approaches can be designed so that the resulting hybrid translations show an improvement over the individual "component" translations. As a first step towards achieving this objective we have developed a parallel corpus with source text and the corresponding translation output from a number of machine translation engines, annotated with metadata information, capturing aspects of the translation process performed by the different MT systems. This corpus aims to serve as a basic resource for further research on whether hybrid machine translation algorithms and system combination techniques can benefit from additional (linguistically motivated, decoding, and runtime) information provided by the different systems involved. In this paper, we describe the annotated corpus we have created. We provide an overview on the component MT systems and the XLIFF-based annotation format we have developed. We also report on first experiments with the ML4HMT corpus data.

Keywords: Machine Translation, System Combination, Annotated Corpus

1. Introduction

Machine translation (MT) is an active field of research with many different paradigms that try to solve the underlying translation problems. In recent years, an important focus for research has been investigating how hybrid MT as well as system combination methods based on the translation output of several translation engines can be designed and implemented so that the resulting translations achieve an improvement over the individual component parts.

One of the main objectives in our research within the META-NET Network of Excellence[1] is a) to provide a systematic investigation and exploration of optimal choices in hybrid machine translation, and b) to support hybrid MT design using sophisticated machine learning (ML) techniques.

As a first step towards achieving these objectives, we have developed a parallel corpus containing source text and the corresponding translation output of several MT systems, representing carefully selected machine translation paradigms, annotated with meta-information, capturing details of the translation process performed by the different MT systems. By including fine-grained and heterogeneous system-specific information as meta-data in the translation output (rather than just providing surface strings), we want to provide rich features for machine learning methods to optimise combination in hybrid MT.

The first release of the ML4HMT corpus is available at `http://dfki.de/ml4hmt/`; it comprises annotated translation output from five MT systems, namely:

- Joshua (Li et al., 2009);
- Lucy (Alonso and Thurmair, 2003);
- Metis (Vandeghinste et al., 2008);
- Apertium (Ramírez-Sánchez et al., 2006);
- MaTrEx (Penkale et al., 2010).

The language pairs inside the corpus are the listed below, all language pairs are available in both directions:

- English ↔ German;
- English ↔ Spanish;
- English ↔ Czech.

In this paper, we describe the training and development data used to generate and compile the corpus (Section 2), the translation engines used for building the corpus (Section 3), and then present its file format in more detail (Section 4). We conclude by giving a summary and an outlook to future work in Section 5.

2. Data

2.1. Annotation Source Data

As a source of the data to be included and annotated in the corpus we decided to use the WMT 2008 (Callison-Burch et al., 2008) "news-test2008" news test set, which is a set of 2,051 sentences from the news domain translated to all languages of our interest, namely English, Spanish, German, and Czech, but also some other languages like French

[1] More information available at `http://meta-net.eu/`

or Hungarian. This test set was provided by the organizers of the Third Workshop on Machine Translation (WMT) in 2008 as test data for the shared translation task. The test set from WMT 2010 (Callison-Burch et al., 2010) has been reserved for final tests in the future (if needed).

2.2. Training Data-Driven Systems

Some of the MT systems used in this work are data-driven (Joshua and MaTrEx). They require a) parallel data for translation phrase pair extraction, b) monolingual data for language modelling, and c) parallel development data for tuning of system parameters. Originally we intended to use the Europarl corpus (Koehn, 2005) for training purposes, but since this widely used parallel corpus did not include Czech at the time when we started the development of the corpus, we decided to make use of the Acquis (Steinberger et al., 2006) and News Commentary parallel corpora[2] instead.

The development data sets required for tuning the SMT systems were taken from the WMT 2008 "nc-test2007" test set[3], which consists of 2,007 sentences from the news-commentary domain and is available in English, French, Spanish, German, and Czech.

3. System Descriptions

3.1. Joshua

Description

Joshua, later referred to as system $t1$, is an open-source toolkit for statistical machine translation, providing a full implementation of state-of-the-art techniques making use of synchronous context-free grammars (SCFGs). The decoding process features algorithms such as chart-parsing, n-gram language model integration, beam-and-cube-pruning, or k-best extraction, while training includes suffix-array grammar extraction and minimum error rate training (MERT).

Annotation

In our meta-data annotations, we provide the output of the decoding process given the "test set", as processed by Joshua (SVN revision 1778). The annotation set contains the globally applied feature weights and for each translated sentence:

- the full output of the produced translation with the highest total score (among the n-best candidates);
- language model scores;
- translation table scores;
- the scores from the derivation of the sentence (phrase scores);
- merging/pruning statistics from the search process.
- each translated sentence, represented by a hierarchical phrase, containing zero or more tokens and pointing to zero or more child phrases;
- word alignment.

3.2. Lucy

Description

The Lucy RBMT system, system $t2$, uses a sophisticated RBMT transfer approach with a long research history. It employs a complex lexicon database and grammars to transform a source sentence into a target language representation. The translation of a sentence is carried out in three major phases: 1) analysis, 2) transfer, and 3) generation.

Annotation

In addition to the translated target text, Lucy provides information about the tree structures that have been created in the three translation phases and which have been used to generate the final translation of the source text. Inside these trees, we can find:

- information about part-of-speech;
- phrases;
- lemma information;
- word/phrase alignment.

In our meta-data annotations, we provide "flattened" representations of the parse trees. For each token, the set of annotated data may contain:

- allomorphs;
- canonical representations;
- linguistic categories;
- the surface string.

3.3. Metis

Description

The Metis system, system $t3$, computes corpus-based translations on the basis of a monolingual target corpus and a bilingual dictionary. The bilingual dictionary functions as a flat translation model that provides n translations for each source word. The most probable translation given the context is then selected by consulting the statistical model built from the target language corpus.

Annotation

Meta-data information for Metis is extracted from the set of final translations ranked by the Metis search engine. For each translation we obtain:

- the score computed during the search process;
- information about part-of-speech;
- lemma information;
- morphological features which are grouped under one feature derived from the source token and may include gender, number, tense, etc.

[2] The WMT News Commentary parallel corpus contains news text and commentaries from the Project Syndicate and is provided as training data for the series of WMT translation shared tasks (See `http://statmt.org/`). Version 10 was released in 2010 and is available in English, French, Spanish, German, and Czech.

[3] Note that this test set is different (in domain, size, and contents) from the aforementioned "news-test2008" test set.

3.4. Apertium

Description

Apertium, system $t4$, originated as one of the machine translation engines in the project OpenTrad, which was funded by the Spanish government. Apertium is a shallow-transfer machine translation system, which uses finite state transducers for all of its lexical transformations, and hidden Markov models for part-of-speech tagging, or word category disambiguation. Constraint grammar taggers are also used for some language pairs (e.g., Breton $\leftrightarrow$ French).

Annotation

We use Apertium version 3.2. Our meta-data annotation includes:

- information about part-of-speech;
- lemma information;
- syntactic information.

For English → Spanish, we have used the following commands: `en-es-chunker` (for syntax information), `en-es-postchunk` (for tags and lemmas), and `en-es` (for the translation).

3.5. MaTrEx

Description

The MaTrEx machine translation system, system $t5$, is a combination-based multi-engine architecture developed at Dublin City University exploiting aspects of both the Example-based Machine Translation (EBMT) and SMT paradigms. The architecture includes various individual systems: phrase-based, example-based, hierarchical phrase-based, and tree-based MT. For the corpus data produced here we used the standard Moses (Koehn et al., 2007) phrase-based SMT system as integrated into MaTrEx.

Annotation

We obtained sentence translations from MaTrEx using its phrase-based SMT system which decomposes the source side to phrases (n-grams), finding their translation and composing them to a target language sentence which has the highest score according the SMT model. Meta-data annotations for each sentence translated by MaTrEx include:

- scores from each model;
- translation probability for each phrase;
- future cost estimate for each phrase.

Also, information about unknown words is included.

4. Corpus Description

We have developed a new dedicated format derived from XLIFF (XML Localisation Interchange File Format) to represent and store the corpus data. XLIFF is an XML-based format created to standardize localisation. It was standardised by OASIS in 2002 and its current specification is v1.2, released on February 1, 2008; see `http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html`.

An XLIFF document is composed of one or more `<file>` elements, each corresponding to an original file or source. Each `<file>` element contains the source of the data to be localised and the corresponding localised (translated) data for one locale only. The localisable texts are stored in `<trans-unit>` elements, each having 1) a `<source>` element to store the source text, and 2) an optional `<target>` element to store the translation.

We defined new elements adding to the basic XLIFF format (inside the `"metanet"` namespace) allowing to store a wide variety of meta-data annotation of the translated texts by different MT systems (tools). The tool information is included in the `<tool>` element appearing in the header of the file. Each tool can have several parameters (i.e. model weights) which are described in the `<metanet:weight>`.

Annotations are stored in the `<alt-trans>` element within the `<trans-unit>` elements. The `<source>` and `<target>` elements in the `<trans-unit>` elements refer to the source sentence and its reference translation, respectively. The `<source>` and `<target>` elements inside `<alt-trans>` elements represent the input and output of a particular MT system (tool). Tool-specific scores assigned to the translated sentences are stored in the `<metanet:scores>` element while the derivation of the translation is specified in the `<metanet:derivation>` element. Its content is tool-specific.

The full format specification is available as an XML schema. An example annotation is shown in Figure 1.

5. Usability of the Corpus

We compared the performance of the contributing systems ($t1$-$t5$) on the sentence level, using two popular metrics, WER and smoothed-BLEU (Lin and Och, 2004). Table 1 shows the percentage of the cases that a specific system gave the best translation for a sentence according the two sentence-level metrics[4].

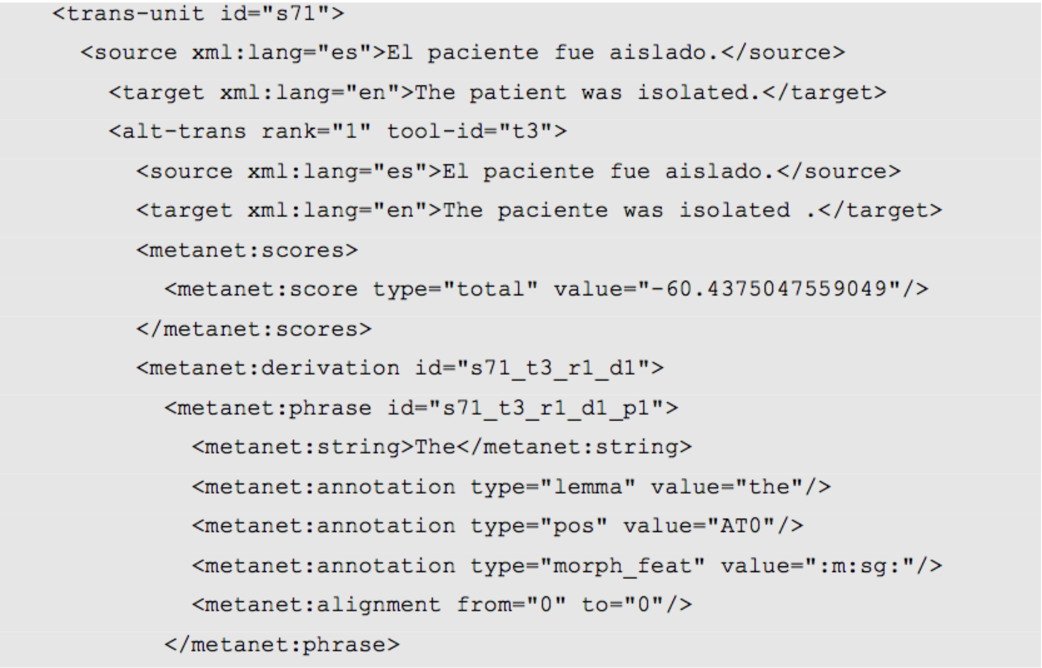
This indicates that the various MT systems included in the corpus perform complementary to each other. We also show the overall BLEU score of each system and compare that to the "optimal" scores achievable if it would be possible to choose the best sentence of each system. This indicates the possibilities of improvement by a sentence-selection approach given the corpus. We believe that even higher performance would be possible using more sophisticated system combination methods.

Additionally, based on the Spanish → English subset of this corpus, Okita and van Genabith (2011) have developed a consensus-based approach with an improvement 2 points BLEU over the single best system on the test set, whereas Avramidis (2011) has presented positive correlation of several features obtained from our corpus. The work presented in Federmann et al. (2011) reports on experiments on a stochastic substitution system based on our corpus data.

6. Conclusion

We have developed an annotated hybrid sample MT corpus which contains a set of 2,051 sentences translated by five

[4] measured over the development set; ties allowed



```
  <trans-unit id="s71">
    <source xml:lang="es">El paciente fue aislado.</source>
      <target xml:lang="en">The patient was isolated.</target>
      <alt-trans rank="1" tool-id="t3">
        <source xml:lang="es">El paciente fue aislado.</source>
        <target xml:lang="en">The paciente was isolated .</target>
        <metanet:scores>
          <metanet:score type="total" value="-60.4375047559049"/>
        </metanet:scores>
        <metanet:derivation id="s71_t3_r1_d1">
          <metanet:phrase id="s71_t3_r1_d1_p1">
            <metanet:string>The</metanet:string>
            <metanet:annotation type="lemma" value="the"/>
            <metanet:annotation type="pos" value="AT0"/>
            <metanet:annotation type="morph_feat" value=":m:sg:"/>
            <metanet:alignment from="0" to="0"/>
          </metanet:phrase>
```

Figure 1: Example of annotation from the ML4HMT corpus.

	Systems				
	$t1$	$t2$	$t3$	$t4$	$t5$
Ranked 1st WER [%]	26.44	37.56	5.85	16.00	58.54
Ranked 1st s-BLEU [%]	14.73	38.63	5.85	23.41	29.95
Overall BLEU	12.80	14.94	8.29	13.34	14.47
Optimal WER	17.62				
Optimal s-BLEU	18.95				

Table 1: Preliminary investigation of the usability of the corpus for Hybrid MT

different MT systems[5] (Joshua, Lucy, Metis, Apertium, and MaTrEx) in six translation directions (Czech → English, German → English, Spanish → English, English → Czech, English → German, and Spanish → English) and annotated with various meta-data information provided by the MT systems.

This resource will be used for feature extraction and training machine learning methods that combine translation output from various MT systems in hybrid MT. The corpus is available from the internet and has already been released to and used by the participants of the Shared Task on Applying Machine Learning techniques to optimising the division of labour in Hybrid MT (ML4HMT-2011)[6].

Acknowledgments

This work has been funded under the Seventh Framework Programme of the European Commission through the T4ME contract (grant agreement no.: 249119) and was supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (`www.cngl.ie`) at Dublin City University. Also, it was partially supported by the Czech Science Foundation (grant no. P103/12/G084) and partially funded by the Spanish Ministry of Economy and Competivity through the Juan de la Cierva fellowship program. The authors would like to thank Felix Sasaki for his support on the XLIFF format. Also, we are grateful to the anonymous reviewers for their valuable feedback and comments.

7. References

Juan A. Alonso and Gregor Thurmair. 2003. The Comprendium Translator System. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA.

Eleftherios Avramidis. 2011. DFKI System Combination with Sentence Ranking at ML4HMT-2011. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November. META-NET.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proceed-*

[5] Not all systems available for all language pairs.

[6] See `http://dfki.de/ml4hmt/`

ings of the Third Workshop on Statistical Machine Translation, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan, editors. 2010. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics, Uppsala, Sweden, July.

Christian Federmann, Yu Chen, Sabine Hunsicker, and Rui Wang. 2011. DFKI System Combination using Syntactic Information at ML4HMT-2011. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November. META-NET.

Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of Annual meeting of the Association for Computation Linguistics (ACL), demonstration session*, pages 177–180, Prague, Czech, June.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit 2005*.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An Open-Source Toolkit for Parsing-Based Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, March. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Tsuyoshi Okita and Josef van Genabith. 2011. DCU Confusion Network-based System Combination for ML4HMT. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November. META-NET.

Sergio Penkale, Rejwanul Haque, Sandipan Dandapat, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. Forcada, and Andy Way. 2010. MaTrEx: the DCU MT system for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 143–148, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, and Mikel L. Forcada. 2006. Opentrad apertium open-source machine translation system: an opportunity for business and research. In *Proceeding of Translating and the Computer 28 Conference*, November.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, and Dan Tufis. 2006. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 2142–2147.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, Denver, Colorado.

Vincent Vandeghinste, Peter Dirix, Ineke Schuurman, Stella Markantonatou, Sokratis Sofianopoulos, Marina Vassiliou, Olga Yannoutsou, Toni Badia, Maite Melero, Gemma Boleda, Michael Carl, and Paul Schmidt. 2008. Evaluation of a machine translation system for low resource languages: METIS-II. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco, May.