АБУМАТРАН
ABUMATRAN

Abu-MaTran

Automatic Building of Machine Translation

PIAP- GA-2012-324414

D3.1c. Acquisition for cycle 3

Dissemination level	Public
Delivery date	2015/12/31
Status and version	Final, v1.0
Authors and affiliation	Miquel Esplà-Gomis (UA), Nikola Ljubešić (UZ), Sergio Ortiz-Rojas (Prompsit), Vassilis Papavassiliou (ILSP), Prokopis Prokopidis (ILSP), Víctor Sánchez-Cartagena (Prompsit) and Antonio Toral (DCU)

Project funded by the European Community under the Seventh Framework Programme for Research and Technological Development

Contents

Executive Summary

This deliverable (D3.1c) describes the activities carried out within the project during the period between the second milestone (or "second development cycle", month 24) and the third milestone period (or "third development cycle", month 36), regarding the acquisition of translation resources to be added to those reported in public deliverable D3.1b (Esplà-Gomis et al., 2014). These resources are needed to build improved statistical and rule-based machine translation (MT) components. The acquired resources are, on the one hand, monolingual and parallel text corpora for statistical MT, harvested from top-level domains (TLD) by using a combination of improved web monolingual and bilingual crawlers developed in the project, and, on the other hand, dictionaries and translation rules, for rule-based MT, using fully automatic techniques for dictionary enrichment and rule learning from parallel corpora, both developed in the project. Work has encompassed six languages: Finnish, Croatian, English, Slovenian, Bosnian, and Serbian. As regards building MT systems, the focus was on the translation between these five languages and English. The work carried out during this period has mainly focused on the news domain. The deliverable refers to papers describing the acquisition techniques and the resulting resources, presented at main conferences and in a mainstream journal.

1 Introduction

This deliverable reports on the activities carried out within the project regarding acquisition during the period between the second milestone (or "second development cycle", month 24) and the third milestone period (or "third development cycle", month 36). The main aim of the deliverable is to report on the automatic acquisition of additional resources to those used in the first and second milestones (cf. D3.1a (Toral et al., 2013) and D3.1b (Esplà-Gomis et al., 2014)) in order to improve the performance of MT. In the case of parallel data crawling, the focus is once more on the Croatian–English language pair; however, in this milestone, four new language pairs have been added: English–Finnish, English–Slovenian, Bosnian–English, and English–Serbian. For the case of acquiring linguistic resources, most of the efforts have focused on Croatian–Serbian.

Section 2 covers the process of crawling monolingual data from top-level domains (TLDs). Section 3 extends this research to the task of crawling parallel data from TLDs. Section 3.2 describes the tool Spidextor, which is aimed at unifying the process of crawling monolingual and parallel data from TLDs by combining two tools: SpiderLing (Suchomel et al., 2012) and Bitextor (Esplà-Gomis and Forcada, 2010). Section 4 discusses about different methods for creating development sets for training general-purpose machine translation (MT) systems. Namely, two options are studied for building these development sets: crowd-sourcing (see Section 4.1) and professional translation (see Section 4.2). The last two sections of this deliverable focus on the creation of resources for rule-based MT; Section 5 covers the acquisition of new words for morphological dictionaries, while Section 6 describes the acquisition of transfer rules from parallel corpora.

2 Monolingual Corpora

2.1 Acquisition of Monolingual Corpora from TLDs

For collecting large monolingual corpora, we used the `Brno` pipeline consisting of a crawler coupled with boilerplate removal, language identification and physical deduplication tools (Suchomel et al., 2012). During the months covered by this deliverable we crawled the TLDs of five countries to get information for five languages: Finnish (Finland, `.fi`), Serbian (Serbia, `.sr`),

Bosnian (Bosnia, .ba), Croatian (Croatia, .hr), and Slovene (Slovenia, .si). While crawling each TLD, beside the main language of the country represented by that domain, we collected English data as well. While the primary language data was used for building language models for statistical MT and for linguistic knowledge acquisition, like constructing morphological lexicons (described in Section 5.2) or extracting multi-word expressions (described in Section 5.3), the bilingual data was used in the task of extracting parallel data (described in Section 3.1).

The amount of data of the final monolingual corpora is the following: 429 million tokens in bsWaC, 1.7 billion tokens in fiWaC (Finnish), 1.9 billion tokens in hrWaC (Croatian), 1.2 billion tokens in slWaC (Slovenian) and 894 million tokens in srWaC (Serbian). Note that, for the corpora, we follow the WaCKy initiative[1] naming convention with the first two letters encoding the language collected and the *WaC* suffix referring to *Web as Corpus*.

3 Parallel Corpora

3.1 Acquisition of Parallel Corpora from TLDs

In the previous milestones Bitextor (Esplà-Gomis and Forcada, 2010) and ILSP Focused Crawler (Papavassiliou et al., 2013) were used to build parallel data from specific websites for which it was known that they contained parallel data (i.e. web pages in L_1 and their translations in L_2). Unfortunately, manually looking for such websites is a slow process, and results in smaller parallel corpora. In order to ease the task of finding websites with parallel content, we propose an approach that combines the TLD language data crawler SpiderLing,[2] which is part of the aforementioned Brno pipeline and the bitext crawlers described in the previous deliverables (i.e. Bitextor and ILSP Focused Crawler). Combining these tools allows us to automatically crawl large amounts of multilingual data by means of SpiderLing from a given TLD, for example English and Finnish from the .fi domain, or English and Croatian from the .hr domain, and then detecting bitexts by using the document alignment modules of the bitext crawlers. In our case, the tool Bitextor was used to perform a shallow search of parallel content in the

[1] http://wacky.sslmit.unibo.it/doku.php?id=start

[2] http://nlp.fi.muni.cz/trac/spiderling/attachment/wiki/WikiStart/spiderling-src-0.77.tar.xz

data crawled with SpiderLing. Using this strategy we obtained the following parallel corpora:

- English–Finnish: The `fienWaC 1.0` corpus (Ljubešić et al., 2016c) was created by crawling a collection of 71,892 websites, 12,183 of them yielding parallel data; it consists of 2,154,652 pairs of segments and 70,038,256 words. A version with more restrictive parameters for document alignment (and therefore, with higher quality) has been published consisting of 890,882 pairs of segments and 28,616,021 words.

- English–Croatian: The `hrenWaC 2.0` corpus (Ljubešić et al., 2016a) was created by crawling a collection of 25,924 websites, 6,228 of which yielded parallel data; it is slightly smaller than the fienWac corpus, and consists of 1,166,732 pairs of segments and 53,581,900 words. The filtered version of this corpus consists of 698,097 pairs of segments and 33,722,383 words.

- English–Serbian: The `srenWaC 1.0` corpus (Ljubešić et al., 2016b) was created in the same manner as the first two corpora. We have produced only the filtered version of this corpus. It consists of 391,953 segments and 21,671,703 words.

This strategy was also used to create the English–Finnish corpus used for the news translation shared task in the Workshop of Statistical Machine Translation 2015 (WMT'15) in which our submission (Rubino et al., 2015) ranked among the first systems competing in this edition of the shared task. Besides using Bitextor, the ILSP Focused Crawler was also used to crawl the $1,000$ largest hotspots of the `.fi` domain, that is, websites with the highest score s, where $s = sum(N_{en}, N_{fi}) * min\{N_{en}, N_{fi}\}/max\{N_{en}, N_{fi}\}$ and N_{en} and N_{fi} stand for the number of acquired English and Finnish web pages per website respectively, and could therefore be considered the most bitext-productive multilingual websites.

From these websites a total of $58,839$ document pairs were identified.[3] Finally, Hunalign[4] (Varga et al., 2007) was applied on these document pairs, resulting in 1.2 million segment pairs after duplicate removal.

[3] $8,936$ document pairs were obtained by their URL similarity, $17,288$ were obtained by their image co-occurrence, and $32,615$ were obtained by their structural similarity.

[4] `http://mokk.bme.hu/en/resources/hunalign/`

3.2 Spidextor

Motivated by the good results obtained by combining SpiderLing and Bitextor for building large parallel corpora, a new software package has been released with the name of Spidextor (as a combination of the names of its two crucial parts – SpiderLing and Bitextor) that includes both tools and provides a collection of scripts for using them together. This new tool is described in (Ljubešić et al., 2016) and is available from GitHub.[5] The aim of this new tool is to enable crawling a TLD for documents written in specific languages, and later on matching documents written in different languages that are probably translations of each other. In order to combine Bitextor and SpiderLing, some adaptations were made on these tools for the versions distributed in this new package. These changes are described in the next sub-sections.

3.2.1 SpiderLing modifications

SpiderLing was primarily built to produce monolingual corpora. Therefore, smaller modifications of the code (20 lines inserted or changed) had to be introduced to enable the user to define multiple languages of interest. Thereby, all documents written in any of the languages are kept in the crawl.

Since SpiderLing uses a simple distance-based language identification procedure (as it was meant to discriminate between documents written in the language of interest and all other languages), having now multiple languages in our crawl, we included in our process one additional run of `langid.py`[6] on the output of SpiderLing to doublecheck the predicted language and filter out or reclassify the wrong predictions.

3.2.2 Bitextor integration

The logic added on top of Bitextor to process the SpiderLing output consists of two scripts: one that transforms the SpiderLing output to the Bitextor's `.lett` format, and another that enables the user to define the language pairs of interest for the task, together with all the paths necessary to run Bitextor, one of which is a small bilingual lexicon which can improve the bitext

[5] `https://github.com/abumatran/spidextor`

[6] `https://github.com/saffsd/langid.py`

extraction results.[7] The second script also produces and runs a makefile in a parallel fashion. All processing by Bitextor occurs at the level of each Internet domain, making the parallelisation of the process straightforward.

4 Development Sets

In the previous two milestones of the project we had already targeted the translation of general-purpose text by evaluating our MT systems on such a corpus: a 1,000-sentence subset of the test set of the WMT'13 shared task,[8] which is made up of news stories that cover different themes (e.g. politics, finance, sport, culture), cf. Section 2.1.1 of the deliverable D4.1b (Forcada et al., 2014) for a related discussion during milestone 2. While this corpus was used as a test set, we did not count with an appropriate general purpose corpus to be used as a development set.

Now in milestone 3 we aim to improve the performance of MT for general purpose texts by acquiring a suitable development set. Namely, we translate a subset (the first 25 news stories, accounting for $1,011$ sentences) of the English side of the test set provided for WMT'12.[9]

We obtain translations of this data set in two ways: professional translation and crowdsourcing. While professional translations lead to a higher quality parallel data set, which should result in a positive impact on the final MT output, its cost can be close to an order of magnitude higher than crowdsourcing.

In this sense, the results obtained by Zbib et al. (2013) are specially interesting, since the authors built MT systems for Arabic–English using development sets that were professionally translated and crowdsourced. They compared tuning with one reference (either professional or crowdsourced) and using both together as multiple references. The best results were obtained using the latter setup, i.e. using both references. When using only one reference, the professional translation led to better results than the crowdsourced one, as would have been expected.

In our experiments we aim to corroborate these results for English–Croatian and also compare the use of professional and crowdsourced trans-

[7] For both of our language pairs we use small bilingual lexicons extracted automatically from phrase tables built on existing parallel data.

[8] `http://www.statmt.org/wmt13/translation-task.html`

[9] `http://www.statmt.org/wmt12/translation-task.html`

lations for tuning. The rest of this section describes the acquisition of the development sets. Description of MT systems using them can be found in Section 2 of Deliverable 4.1c. Finally, results and analyses of those systems are provided in Section 2 of Deliverable 5.1c.

4.1 Crowdsourcing

Crowdsourced translations were obtained with the CrowdFlower platform.[10] CrowdFlower provides a cheap and fast method for collecting annotations from a broad base of paid non-expert contributors over the Web. It works in a similar way to Amazon's Mechanical Turk (Snow et al., 2008).

In the task we defined, the contributors had to translate manually, from English into Croatian, the development set (1,011 sentences), which was divided in working tasks of 10 sentences each. We requested two judgments per source sentence in order to obtain two reference translations. Pay was set to 80 cents of dollar per task (i.e. 8 dollar cents per sentence), orders of magnitude below the market price of professional translation. The total cost, including Crowdflower's fee (20%), was 196 dollars.

Some instructions were provided to guide contributors. Mainly, they were asked to produce the translations manually and specifically told not to use MT. The instructions, as provided to the contributors, follow:

```
Each job contains 10 sentences from newstories in English.
Your task is to translate them into Croatian. Note that:
- The news are to be manually translated.
  The use of machine translation is NOT allowed
- Failure to follow these instructions will discard you as a worker!
```

This crowdsourcing platform permits configuring the jobs using a number of options. We used some of them with the aim of obtaining translations of a reasonable quality. The options that we used follow:

- *Geography*: One can select a set of countries from which workers are allowed to work on the job. We limited this to Croatia.
- *Performance level*: Contributors of the platform fall into three levels, according to their performance. Our jobs were limited to contributors in the top level, defined by Crowdflower as "the highest performance

[10] `http://crowdflower.com/`

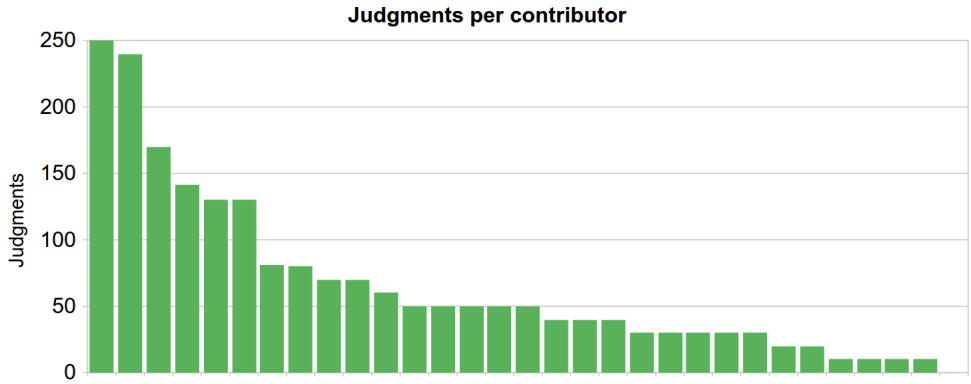

Figure 1: Distribution of the total number of judgments (y axis) per contributor (x axis) in crowdsourced translations of the development set from English to Croatian.

contributors who account for 7% of monthly judgments and maintain the highest level of accuracy across an even larger spectrum of CrowdFlower jobs [compared to contributors in levels 1 and 2]".

- *Speed trap*: If set, contributors are automatically removed from the job if they take less than a specified amount of time to complete a task. The time trap was set to 300 seconds and, as previously mentioned, our jobs contained tasks of 10 translations each. Hence, if translating a sentence takes a worker less than 30 seconds on average, the worker is automatically removed from the job.

- *Maximum number of judgments per contributor*: This is the total number of judgments that any one contributor can complete. We set this to 250 so any contributor could translate at most around 25% of the development set.

Thirty different contributors participated in the task. The distribution of the total number of judgments per contributor is shown in Figure 1. Contributors had the option of evaluating the task by providing scores (from 1 to 5, where 1 is the worst score and 5 is the best one) associated to different aspects: overall judgment, clarity of the instructions, ease of the job and pay. Seventeen contributors (57%) did so, and ranked most aspects above 4 (4.6 instructions, 4.2 pay, 4.5 overall) except for the ease of the job (3.8).

To verify the quality of the translations, a sample of 50 sentences per contributor were evaluated by 2 native speakers. Translations were found to be generally acceptable but far from professional quality.

4.2 Professional Translations

We requested quotes to translate the development set to a number of language service providers that work with our language direction and hired the one that provided the lowest quote (rate of 0.05 euros per word leading to a total of 930 euro for our dataset of over 18,000 words).

Compared to the cost of a single reference translation by means of crowdsourcing (98 dollars), professional translation results roughly one order of magnitude more expensive. Timewise, professional translation took 1 week while the two reference translations obtained by means of crowdsourcing were completed in less than a day. It should be noted, however, that the professional translation was completed by 1 translator while the crowdsourced translations were carried out by 30 contributors (cf. Figure 1).

Table 1 shows several quantitative statistics of the development sets produced both by professional and crowdsourcing translators. Namely, we show the number of words in the translation (column # words), the average sentence length (avg sent length), the vocabulary size (voc size) and the type to token ratio (TTR).[11]

Development set	**# words**	**avg sent length**	**voc size**	**TTR**
Professional	19,586	19.37	6,063	0.3096
Crowd 1	18,642	18.44	6,313	0.3386
Crowd 2	18,513	18.31	6,315	0.3411

Table 1: Quantitative statistics of the translated development sets

Sentences in the professional translation are 5 to 5.8% longer (depending to which crowdsourced reference we compare it). This corroborates the findings by Zbib et al. (2013), where crowdsourced translations (from Arabic into English) were 10% shorter than professional ones.

[11] TTR of a text is the ratio obtained by dividing the types (i.e. total number of different words in the text) by its tokens (i.e. total number of words that occur in the text). A high TTR indicates a high degree of lexical variation while a low TTR indicates the opposite.

Despite being shorter, crowdsourced translations contain a slightly higher number of unique words (column vocabulary size), i.e. showing more variability in the words used. This is probably a consequence of the fact that these translations were produced by 30 contributors and thus they might be less consistent than the professional translation, done by one translator.

5 Dictionaries

5.1 Acquisition of Bilingual Dictionaries from Comparable Corpora

This work was completed during a secondment from Prompsit to UA. The aim was to build a software that could benefit of Wikipedia as a source of comparable texts. We chose the Slovenian and Croatian Wikipedias because both are languages of interest for the project and both are close enough, so that we could expect that the corresponding Wikipedias could present equivalent contents for articles related, and that, therefore, common terms being mutual translation could be extracted.

5.1.1 Data used

The data used came from Wikipedia dumps processed to extract raw text, preserving some markup for titles of articles, and sections and subsections. All wikipedia articles were preserved as a single text document. In this big document, each Wikipedia article starts with the main title and ends with the next Wikipedia article (the next title).

Document matchings were established using Wikipedia's interlanguage links that normally are displayed along with the articles to read about the same concept in other language. English Wikipedia links were used as a pivot because we found that they were more complete.

Considering articles with any matching in the other language, only $137,015$ Slovenian and $126,347$ Croatian articles were preserved for the task, for a total of 32 million and 34 million words, respectively.

5.1.2 Statistical approach

The idea was to use general SMT software to count word co-occurrence, and then some metric as edit distance to try to enforce predictions of good new

entries for the system.

Each article was synthetically rewritten in one line, so that it would be treated as a single segment. All stopwords, considered as any word with a frequency below 75, were removed from articles. The GIZA++ toolkit (Och and Ney, 2003) was used to discover "alignments" (co-occurences or sequences of occurences).

The Jaro-Winkler distance (Winkler, 1990) was intended to give a confidence to cognates found in each alignment (after testing Levenshtein distance (Wagner and Fischer, 1974; Levenshtein, 1966) and Jaccard index (Jaccard, 1912) as well). Results were very discouraging. Performing a manual evaluation of a random sample of 100 results, only 9 percent of retrieved terms were valid entries. No general rule was evident to filter out bad entries from good entries.

5.1.3 Word embeddings

After failing in this approach, and trying still to take advantage of the same Wikipedia data, we tried to implement the method by Mikolov et al. (2013) by word2vec software for word embeddings from the same authors to reproduce the claim that word2vec organises independently words of non-aligned, non-paired texts in several languages in a way that a linear transform (a dot product of a matrix) could relate accurately elements from one vector space, the embeddings of one of the languages, to the other.

As pointed in the article, a stochastic gradient descent algorithm was used to estimate that matrix. A total of 10,000 entries coming from Apertium dictionaries were used as anchors (training examples) to perform the estimate.

Our results were far from those obtained by Mikolov et al. (2013). Most probably this is due to the fact that the size of our dataset is not that big (it has been shown that for word2vec to obtain optimal results it needs vast amounts of data)·

5.2 Morphological lexicons

We have developed a technique (Ljubešić et al., 2015b) for extending already existing morphological lexicons in Apertium by supervised machine learning. The task is framed as a ranking task: a collection of candidates (lemma, paradigm) are ranked for human inspection for a given word that is to be

included in a morphological lexicon lexicon. This ranking is provided by a statistical model that builds on the following features:

- stem features – capture information about the stem obtained from the surface form after removing the suffix candidate
- lexicon features – represent the information from the existing lexicon about the relation between a paradigm and suffixes and prefixes of stems and lemmata that belong to that paradigm
- corpus features – they are extracted from a large corpus of the language in question, and they contain information such as whether a candidate lemma was seen in the corpus, or the ratio of forms of the whole paradigm attested in the corpus.

The ranking is evaluated by using the mean reciprocal rank (MRR) metric. In this case, we obtain a MRR of 0.83 for Croatian, which means that, for 77% of unknown words, a correct candidate (lemma,paradigm) can be found in the top position of the ranking, while in 95% of cases, a correct candidate (lemma, paradigm) can be found in top five positions.

5.3 Multiword units

We have developed a tool – DepMWEx[12] – that enables writing patterns at the level of dependency syntax (and lower levels of abstraction) for extracting multiword expression candidates from parse trees. The extracted candidates are later weighted by standard association measures like logDice, the logarithm of the Sørensen–Dice coefficient.[13]

By writing the required patterns and applying them on the parsed Croatian and Serbian web corpora, we automatically produced MWE lexicons for both languages (Ljubešić et al., 2015a). Both the lexicon for Croatian[14] and for Serbian[15] were made available under the CC-BY-SA license.[16]

[12] https://github.com/nljubesi/depmwex

[13] https://en.wikipedia.org/wiki/S\%C3\%B8rensen\%E2\%80\%93Dice_coefficient

[14] http://nlp.ffzg.hr/resources/lexicons/hrmwelex/

[15] http://nlp.ffzg.hr/resources/lexicons/srmwelex/

[16] https://creativecommons.org/licenses/by-sa/4.0/

The Croatian resource consists of 46,293 head lexemes and 12,750,029 MWE candidates while the Serbian one consists of 23,594 head lexemes and 3,279,864 MWE candidates. The precision of the extraction process is around 50%.

6 Rules

During the last year, the rule inference algorithm described in the previous report (Sánchez-Cartagena et al., 2015) was used to create a new rule-based MT system building on Apertium for the Serbian–Croatian language pair. Since there was no MT system already developed for this language pair and the amount of parallel corpora available is relatively small, building a new rule-based MT system with the help of the rule inference approach is the most cost-efficient strategy. Moreover, since Serbian and Croatian share a substantial part of the vocabulary, the human effort invested in the creation of the bilingual lexicon was small.

The algorithm by Sánchez-Cartagena et al. (2015) produced a set of high-quality shallow-transfer from a small piece of the SETimes parallel corpus. These rules encoded some of the most important grammatical differences between Serbian and Croatian and clearly outperformed a rule-based MT system with no rules at all. They were edited in order to produce the final version of the Apertium Serbian–Croatian MT system.

The current version of the Serbian–Croatian MT system achieves a BLEU score of 0.805 and a TER score of 0.093 on a 350-segment test set. The leave-as-is baseline (using the rules as inferred)achieves a BLEU score of 0.737 and a TER score of 0.125. However, Google Translate outperforms our rule-based system with a BLEU score of 0.823 and a TER score of 0.087.

As pointed out in deliverable D3.1b (Esplà-Gomis et al., 2014), the algorithm by Sánchez-Cartagena et al. (2015) can be used, together with existing dictionaries, to build a hybrid MT system that consists of an SMT system whose translation model is enriched with the existing dictionaries and a set of transfer rules inferred from the training parallel corpus. However, the first evaluations carried out (Sánchez-Cartagena et al., 2014) showed no improvement over an SMT system trained on the same data.

During the last year, this hybrid strategy has been studied in depth (Sánchez-Cartagena et al., 2016). The conclusion was that, by taking advantage of how the linguistic resources are used by the RBMT system to segment the

source-language sentences to be translated, a significantly greater translation quality than that of a baseline SMT system can be achieved. Moreover, the translation quality achieved by the hybrid system built with automatically inferred rules is similar to that obtained by those built with hand-crafted rules.

7 Conclusion

This deliverable has covered the work done in the area of acquisition (work package WP3) during the period of the third milestone of the project (M25–M36). We have worked on the acquisition of two types of resources: corpora and linguistic data. These two types of resources can be loosely identified with the different MT paradigms that will make use of them (statistical and rule-based, respectively).

As regards the automatic acquisition of corpora, the previous work done so far in this work package had focused on a single language pair: English–Croatian. In this deliverable, four new language pairs have been added: English–Finnish, English–Slovenian, Bosnian–English, and English–Serbian. The acquisition of textual data has been covered, both for monolingual and parallel data from TLDs. Crawling TLDs directly improves the degree of automatism of the process, since it is not necessary to look for websites containing parallel data. In fact, a strategy has been defined for combining the tool SpiderLing and Bitextor in order to crawl monolingual data from a TLD and, in parallel, looking for parallel data, in an only pipeline. This strategy is implemented in the tool Spidextor.

Moving on to linguistic resources, we have worked on the acquisition of dictionaries and rules. On the one hand, some experiments have been done in order to try to extract word-translations from comparable corpora. Unfortunately, the strategies tried did not produce results of the quality expected. In this line, a new method was defined for detecting the most suitable paradigm in a morphological dictionary for a given word form using a statistical approach. Finally, a method for extracting multi-word expressions from monolingual corpora was defined, which produced an acceptable recall, although the accuracy is still low (about 50%).

The final part of the work carried out during this period is aimed at applying the method by Sánchez-Cartagena et al. (2015) for inferring transfer rules for rule-based MT for the pair of languages Croatian–Serbian. The

results obtained for this objective confirmed that this strategy was able to improve the performance of the previous version of the Croatian–Serbian Apertium system noticeably, even though it was still slightly lower than other commercial MT systems, such as Google Translate, which use much larger resources.

References

Miquel Esplà-Gomis, Mikel L. Forcada, Nikola Ljubešić, Vassilis Papavassiliou, Prokopis Prokopidis, Sergio Ortiz Rojas, Pirinen Tommi, Raphaël Rubino, and Antonio Toral. Deliverable D3.1b acquisition for cycle 2. Technical report, Abu-MaTran, MSCA-IAPP project PIAP-GA-2012-324414, 2014.

Miquel Esplà-Gomis and Mikel L. Forcada. Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, (93): 77–86, 2010. ISSN: 0032-6585.

Mikel L. Forcada, Sergio Ortiz-Rojas, Tommi Pirinen, Raphaël Rubino, and Antonio Toral. Deliverable D4.1b MT systems for the second development cycle. Technical report, Abu-MaTran, MSCA-IAPP project PIAP-GA-2012-324414, 2014.

Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710, 1966.

Nikola Ljubešić, Kaja Dobrovoljc, and Darja Fišer. *MWELex – MWE Lexica of Croatian, Slovene and Serbian Extracted from Parsed Corpora. *Informatica*, 39(3):293–300, 2015a.

Nikola Ljubešić, Miquel Esplà-Gomis, Filip Klubička, and Nives Mikelić Preradović. Predicting Inflectional Paradigms and Lemmata of Unknown Words for Semi-automatic Expansion of Morphological Lexicons. In *Proceedings of Recent Advances in Natural Language Processing*, pages 379–387, 2015b.

Nikola Ljubešić, Miquel Esplà-Gomis, Sergio Ortiz Rojas, Filip Klubička, and Antonio Toral. Croatian–English parallel corpus hrenWaC 2.0, 2016a. URL http://hdl.handle.net/11356/1058. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Miquel Esplà-Gomis, Sergio Ortiz Rojas, Filip Klubička, and Antonio Toral. Serbian–English parallel corpus srenWaC 1.0, 2016b. URL http://hdl.handle.net/11356/1059. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Miquel Esplà-Gomis, Sergio Ortiz Rojas, Filip Klubička, and Antonio Toral. Finnish–English parallel corpus fienWaC 1.0, 2016c. URL http://hdl.handle.net/11356/1060. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Miquel Esplà-Gomis, Antonio Toral, Sergio Ortiz Rojas, and Filip Klubička. Producing monolingual and parallel web corpora at the same time - spiderling and bitextor's love affair. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013. URL http://arxiv.org/abs/1309.4168.

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

Vassilis Papavassiliou, Prokopis Prokopidis, and Gregor Thurmair. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2506.

Raphael Rubino, Tommi Pirinen, Miquel Esplà-Gomis, Nikola Ljubešić, Sergio Ortiz Rojas, Vassilis Papavassiliou, Prokopis Prokopidis, and Antonio Toral. Abu-Matran at WMT 2015 translation task: Morphological segmentation and web crawling. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 184–191, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-3022`.

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortíz, and Felipe Sánchez-Martínez. The UA-Prompsit hybrid machine translation system for the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 178–185, Baltimore, MD, USA, June 2014.

Víctor M. Sánchez-Cartagena, Juan A. Pérez-Ortiz, and Felipe Sánchez-Martínez. A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. *Computer Speech & Language*, 32(1):46–90, 2015.

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. Integrating rules and dictionaries from shallow-transfer machine translation into phrase-based statistical machine translation. *Journal of Artificial Intelligence Research*, 55:17–61, January 2016.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, 2008.

Vít Suchomel, Jan Pomikálek, et al. Efficient web crawling for large text corpora. In *Proceedings of the Seventh Web as Corpus Workshop (WAC7)*, pages 39–43, 2012.

Antonio Toral, Santiago Cortés-Vaíllo, Gema Ramírez-Sánchez, and Nikola Ljubešić. Abu-matran deliverable D3.1a: Acquisition for the first development cycle. Technical report, Abu-MaTran, MSCA-IAPP project PIAP-GA-2012-324414, 2013.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. *Ams-*

terdam Studies in the Theory and History of Linguistic Science Series 4, 292:247–258, 2007.

Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, January 1974. ISSN 0004-5411. doi: 10.1145/321796.321811. URL `http://doi.acm.org/10.1145/321796.321811`.

William E. Winkler. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359, 1990.

Rabih Zbib, Gretchen Markiewicz, Spyros Matsoukas, Richard M. Schwartz, and John Makhoul. Systematic comparison of professional and crowdsourced reference translations for machine translation. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 612–616. The Association for Computational Linguistics, 2013. URL `http://aclweb.org/anthology/N/N13/N13-1069.pdf`.