АБУМАТРАН
ABUMATRAN

Abu-MaTran

Automatic Building of Machine Translation

PIAP- GA-2012-324414

D3.1d. Acquisition for cycle 4

Dissemination level	Public
Delivery date	2016/12/31
Status and version	Final, v1.0
Authors and affiliation	Víctor Sánchez-Cartagena (Prompsit), Meghan Dowling (DCU), Miquel Esplà-Gomis (UA), Mikel Forcada (UA), Nikola Ljubešić (UZ), Vassilis Papavassiliou (ILSP), Prokopis Prokopidis (ILSP) 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 (D3.1d) describes the activities carried out within the project during the period between the third milestone (or "third development cycle", month 36) and the fourth milestone period (or "fourth development cycle", month 48), regarding the acquisition of translation resources to be added to those reported in public deliverable D3.1c (Esplà-Gomis et al., 2015). As the fourth development cycle is the last one of the project, focus has been put on adapting the approaches described in previous deliverables to new languages and on disseminating the tools and methods already developed by means of software releases and participation in shared tasks. Crawling of Internet top-level domains has been extended to English–Irish while semi-automatic and automatic methods for building inflectional dictionaries have been tested on 5 new languages. The automatic rule inference algorithm used in previous deliverables has been released as an open-source tool and the results of the participation of project partners in different shared tasks highlight the effectiveness of the web crawling technology developed during the project.

1 Introduction

This deliverable reports on the activities carried out within the project regarding acquisition the period between the third milestone (or "third development cycle", month 36) and the fourth milestone (or "fourth development cycle", month 48), regarding the acquisition of translation resources to be added to those reported for previous milestones. As the fourth development cycle is the last one of the project, focus has been put on adapting the approaches described in previous deliverables to new languages and on disseminating the tools and methods already developed by means of software releases and participation in shared tasks. As regards the acquisition corpora (Section 2), crawling of Internet top-level domains has been extended to English–Irish, additional parallel corpora have been acquired for use, and the usage-rights problems relating the distribution of crawled corpora have been addressed. Semi-automatic and automatic methods for building inflectional dictionaries (Section 3) have been tested on 5 new languages. As regards rule learning, Section 4 describes the release of the automatic rule inference algorithm used in previous deliverables as an open-source tool, and a method to weight rules when their application is ambiguous. Finally (Section 5), the results of the participation of project partners in two and the participation of Abu-Matran in two shared tasks —a document alignment shared task and a translation memory cleaning task— highlight the effectiveness of the web crawling technology developed during the project.

2 Corpora

2.1 Acquisition of Parallel Corpora

2.1.1 English–Irish Corpus from the .ie Top Level Domain

The task of acquiring parallel data from Internet top-level domains (TLDs), described in deliverable D.3.1c (Esplà-Gomis et al., 2015), continued during this year. We addressed the Republic of Ireland (`.ie`) TLD for obtaining English–Irish parallel data.

There is a very important difference between the `.ie` TLD and those crawled in the previous years of the project: most of the `.ie` websites are written only in English. However, we are only interested in those sites written in both English and Irish. Downloading all the English monolingual websites

is not necessary, as there is enough English monolingual data available and downloading them would take a long time.[1]

In the TLDs crawled previously in the project with the tool Spidextor (`.fi`, `.hr`), monolingual data from the majority language of the TLD (Finnish and Croatian, respectively) was useful for building MT systems. In this case, however, we need to discard those websites written only in English. Thus, instead of using Spidextor, we introduced a set of new features into Bitextor that allow us to avoid downloading large websites written only in English, namely:

- Support for crawling multiple websites at once.
- Support for following links between different websites.
- Early stopping: if after downloading a certain number of documents from a website, there are no documents for either of the languages of the pair, the website is discarded.

The version we created[2] can now be used as a standalone tool for obtaining parallel data from TLDs.

We run the TLD crawling for 6 weeks. We started the process with a seed list of websites we knew in advance that they contained pages in both English and Irish and let Bitextor follow links to other web sites. After removing nearly-duplicated documents (web pages), we obtained 111 663 documents, from which only 1 079 were written in Irish. We finally obtained a very small parallel corpus with 1 628 segments.

Results confirm our previous observation about the scarcity of websites written in Irish and suggest that longer crawling time is needed despite the adaptation of our crawling tool to the `.ie` TLD.

2.1.2 Parallel Global Voices

ILSP created Parallel Global Voices (PGV), a collection of multilingual corpora with citizen media stories. PGV is a set of parallel and monolingual corpora generated from the Global Voices multilingual group of websites (`https://www.globalvoices.org/`), where volunteers have been publishing

[1] According to `https://www.iedr.ie/`, there are more than 200 000 registered `.ie` domains.

[2] Available at: `http://svn.code.sf.net/p/bitextor/code/branches/tld`

and translating news stories in more than 40 languages since 2004. Prokopidis et al. (2016) report on how this content was crawled and processed to generate resources including 302,600 document pairs and 8,360,000 segment alignments in 756 language pairs. For some language pairs, the segment alignments in this resource are the first open examples of their kind.[3] The authors also evaluated the document alignment methods of the ILSP Focused Crawler, one of the parallel crawlers used and enhanced in the project, in the task of reconstructing the English–Greek parallel collection of the PGV dataset without taking into account the translation links connecting EN and EL documents. The evaluation results showed that on document pairs of non-trivial length ($>$ 500 total tokens in the main content of both documents in a pair), the combination of the document alignment methods of ILSP-FC reached a 97.26% F-score.

2.2 Deferred Crawling to Circumvent Legal Issues of Distributing Crawled Corpora

Project Abu-MaTran has extensively dealt with the use of sentence-aligned web-crawled parallel text or *bitext* to train statistical and neural machine translation systems or to adapt them to a new domain, and has devoted considerable effort to developing bitext crawling techniques and software. While third parties could use the software released by Abu-MaTran to crawl their own corpora, public distribution of web-crawled sentence-aligned bitext sets could be beneficial to them, and indeed, many bitext corpora are actually available in the Internet, with perhaps the most important example being the Europarl corpus.[4] Contrary to what is commonly believed, distribution of web-crawled text is far from being free from legal implications (Tsiavos et al., 2014; Arranz et al., 2013), and may sometimes actually violate the usage restrictions (for instance, according to the Berne Convention, anything that is published without an explicit copyright notice reserves the author all possible rights). As the distribution and availability of sentence-aligned bitext is key to the development of statistical machine translation systems, a paper by researchers at partner Universitat d'Alacant (Forcada et al., 2016) proposes an alternative: instead of copying and distributing copies of web content in the form of sentence-aligned bitext, one could distribute a legally safer *stand-*

[3]PGV is freely available from `http://nlp.ilsp.gr/pgv/`

[4]`http://www.statmt.org/europarl/`

off annotation of web content, that is, sets of files that identify where the aligned sentences are, so that end users can use this annotation to privately and efficiently recrawl the bitexts. The paper describes and discusses the legal and technical aspects of this proposal, and outlines an implementation, which has not yet however been incorporated into the crawling software released by the project.

3 Dictionaries

This section describes the work carried out to develop methods that ease the task of creating new inflectional dictionaries. These dictionaries encode entries as pairs of stems and paradigms, being the stems the static part of the words, i.e. the part that does not change, and the paradigm the collection of suffixes that can be combined to the stem to obtain the different inflected surface forms related to it. Paradigms also relate inflection to morphological information, such as the lexical category, person, number, etc., of the inflected surface form. These dictionaries allow to efficiently store linguistic information that can be later used in a number of applications, such as rule-based machine translation (MT) systems.

When building inflectional dictionaries, it is usual to first define the collection of paradigms, and then start adding new dictionary entries. Adding new entries to these dictionaries is a complex task that requires a certain degree of expertise and knowledge about the morphology of the language involved. The objective of the work carried out is to make it easier to extend inflectional dictionaries following two (complementary) approaches: one aimed at helping non-expert users to manually insert new entries into the dictionary (cf. Section 3.1), and one capable of automatically picking the most promising entry that can then be supervised by an expert (cf. Section 3.2).

3.1 Assisted Building of Inflectional Dictionaries

Partner UA delved into their previous works for developing a method for assisting users to extend monolingual inflectional dictionaries (Sánchez-Cartagena et al., 2012; Esplà-Gomis et al., 2014). The method proposed by Esplà-Gomis et al. (2016) tries to elicitate the knowledge of non-expert users by asking them about the validity of some inflected surface forms from different combinations of stems and paradigms *guessed* automatically from a single surface

form to be inserted in the dictionary. The method proposed is aimed at minimizing the number of forms proposed to the user in order to make the task as easy and simple as possible. The approach is evaluated both with an oracle approach and with human evaluation and involves new language pairs: Maltese, Basque, Catalan, and Spanish. The results confirm the usefulness of the approach proposed, allowing to incorporate non-expert knowledge to the task of building new inflectional dictionaries.

3.2 Semi-Automatic Extension of Inflectional Dictionaries

Partner UZ used the method described in (Ljubešić et al., 2015) and reported in (Esplà-Gomis et al., 2015) to heavily extend existing Apertium inflectional morphological dictionary of Croatian and Serbian (Ljubešić et al., 2016).

At the beginning of the extension process the dictionary consisted of 10,183 entries that were assigned to 413 open-class paradigms, while the final lexicon is 105,358 entries strong with 1,227 open-class paradigms.

The extension process was run in multiple iterations. Each iteration started with identifying surface forms not covered in the dictionary that are most frequent in the Croatian and Serbian web corpora *hrWaC* and *srWaC* (Ljubešić and Klubička, 2014). In the following step for the most frequent out-of-vocabulary forms the (lemma, paradigm) pair candidates were calculated and ranked as described in (Ljubešić et al., 2015). The rankings were then loaded into the web interface presented in Figure 1 which was used by language experts to annotate the correct (lemma, paradigm) pair. At the end of each iteration the annotations were loaded into the dictionary and surface forms from web corpora not covered by the dictionary were recalculated.

The extended Apertium lexicon was used to produce two new easy-to-use inflectional lexicons: the inflectional lexicon of Croatian hrLex (Ljubešić et al., 2016a) and the inflectional lexicon of Serbian srLex (Ljubešić et al., 2016b). Both lexicons consist of (word, lemma, morphosyntactic description) triples enriched with frequency and per-million frequency calculated from the corresponding web corpora.

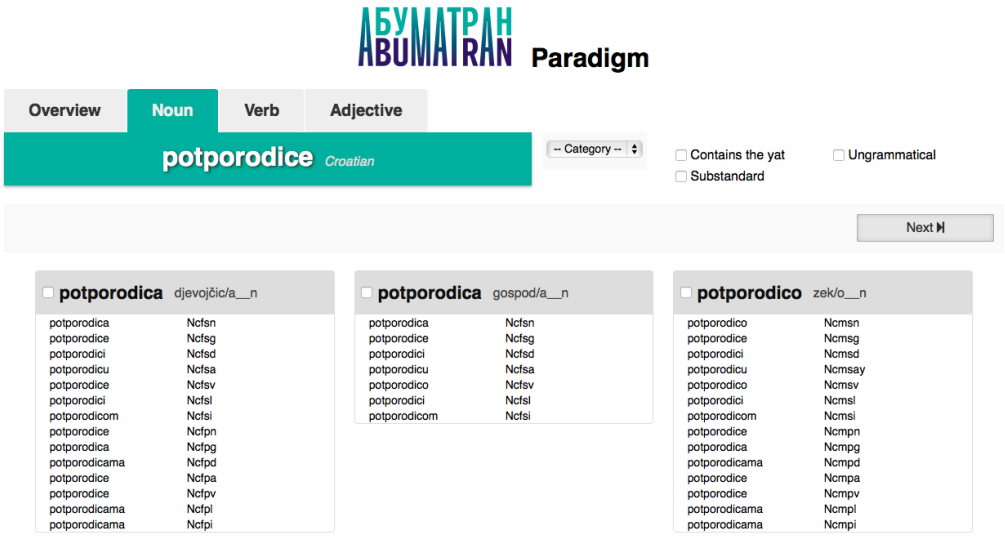

Figure 1: Web interface used for extending the Croatian and Serbian dictionary.

4 Rules

This section describes the work carried out to develop methods for automatically inferring linguistic resources to be used in the structural transfer step of rule-based machine translation systems. Work was focused on releasing a tool that implements the rule inference algorithm previously developed in this project (cf. Section 4.1), and on automatically inferring weights that can be used to choose among ambiguous hand-written rules (cf. Section 4.2).

4.1 Release of rule inference tool

During the last year, a tool (Sánchez-Cartagena et al., 2016) that implements the rule inference algorithm that was used to build the Serbian–Croatian rule-based machine translation system described in deliverable D.3.1c (Esplà-Gomis et al., 2015) has been released.

Recall that, as explained in deliverable D.3.1b (Esplà-Gomis et al., 2014), this rule inference algorithm can make rule-based machine translation a very appealing alternative for under-resourced language pairs because it avoids the need for human experts to handcraft transfer rules and requires, in contrast

to statistical machine translation, a small amount of parallel corpora (a few hundred parallel sentences proved to be sufficient). It is able to produce rules whose translation quality is similar to that obtained by using hand-crafted rules

ruLearn, which is the name of the tool we released, generates rules that are ready for their use in the Apertium platform, although they can be easily adapted to other platforms. The ruLearn source code can be downloaded from `https://svn.code.sf.net/p/apertium/svn/trunk/ruLearn`. It is licensed under GNU General Public License version 3,[5] and distributed as a GNU Autotools[6] package. It currently can only be compiled and executed under GNU/Linux.

ruLearn is written in `Bash` and `Python`. There is an independent command-line program for each step of the algorithm and a wrapper program that executes all the steps. Communication between the different modules is done by writing and reading intermediate files, which makes debugging easier.

4.2 Google Summer of Code: weighted transfer rules

Apertium (Forcada et al., 2011) transfer rules are combinations of fixed-length patterns and actions. When two or more rules match the same segment of the input, there is a *conflict* that needs to be solved. Conflicts are usually *solved* in Apertium in two ways:

- If two patterns of different length match the same segment of the input, the longest pattern is chosen.
- Among matching patterns of the same length (from now on, ambiguous patterns), the first one to appear in the rules file is chosen.

During last summer, a Google Summer of Code[7] project aimed at implementing a more powerful method for deciding among ambiguous patterns was carried out by Russian student Nikita Medyankin under the mentoring of AbuMatran member Víctor M. Sánchez-Cartagena.

[5] `https://www.gnu.org/licenses/gpl-3.0.txt`

[6] `http://www.gnu.org/software/autoconf/` and `http://www.gnu.org/software/automake/`

[7] Google Summer of Code is a program where post-secondary students age 18 and older are paid a stipend to spend their summer break writing code and learning about free/open-source development in one of the organizations selected for the programme. Project Apertium was one such organization.

Apertium transfer module was modified so as to choose the pattern from an ambiguous group with the help of a weights file. In this way, when more than one transfer rule with the same length matches a segment from the input sentence, the rule with the maximum weight is executed. Weights do not depend only on the transfer rule, but also on the particular words of the segment of the input sentence that matched the rule.

Moreover, an algorithm for automatically learning the weights from a parallel corpus was implemented and released. If there is no parallel corpora available for a given language pair, an alternative algorithm that needs only a monolingual corpus in the source language and a language model in target language was also released.

With the implementation of these changes, Apertium linguistic developers can focus on dealing with the general grammatical differences between the languages when writing the rules, and exceptions that affect only some particular words can be automatically inferred. A prime example of the usefulness of this approach is the translation of the construction `noun1` + *de* + `noun2` from Spanish to English. Depending on the particular nouns, the most appropriate translation could be `noun2` + `noun1`, `noun1` + *of* + `noun2`, or `noun2` + *'s* + `noun1`. Manually coding which of the three translations is the best one for each pair of nouns in Spanish would take too much time.

Preliminary experiments carried out after adding some ambiguous rules to the Apertium Spanish–English language pair showed small improvements in BLEU both when the weights were learnt from monolingual and bilingual corpora.

Complete information about the project can be found at `http://wiki.apertium.org/wiki/Weighted_transfer_rules_at_GSoC_2016`.

5 Shared Tasks

This year we have taken part in two shared tasks that concern the processing of parallel data. Specifically they concern alignment and cleaning. We describe our participation in Sections 5.1 and 5.2, respectively.

5.1 Document Alignment Shared Task

Given that Bitextor and ILSP examine the whole content of a webdomain (the former as site copier and the latter as site harvester) with the purpose

of identifying document pairs and finally segment pairs, the modules of pair detection of both tools participated in the WMT 2016 Bilingual Document Alignment shared task of the ACL 2016 First Conference on Machine Translation (Buck and Koehn, 2016).[8] The task was to identify pairs of English and French documents from a given collection of documents such that one document is the translation of the other. The evaluation metric of the task was recall of the known pairs, i.e. what percentage of the aligned pages in the test set were found in a submission. Given that a number of participants pointed out that some predicted document pairs were unfairly counted as wrong, even if their content differed only insignificantly from the gold standard, the organizers also included a soft scoring metric which counts such near-matches as correct.

Papavassiliou et al. (2016) describe a system submitted by ILSP, where several document and collection-aware features were explored in the context of the task. On the test dataset, the submission achieved a recall of 84.93%, even though it does not make use of any language-specific resources like bilingual lexica or MT output. Instead, the system was based on shallow features (including links to documents in the same webdomain, URLs, digits, image filenames and HTML structure) that can be easily extracted from web documents. When de-duplication issues in the test dataset were properly addressed, the system reached a significantly higher recall: in the Soft Scoring Results of the task, the ILSP submission reached a 91.0% recall and was ranked 7th.

For their part, Esplà-Gomis et al. (2016) submitted two systems to the task: one based on Bitextor 4.1 (the last release of the tool at the moment of the task) and a system based on the new version of Bitextor (version 5.0)[9]. Version 5.0 of Bitextor clearly outperformed the previous one, reaching 95% precision and 87.5% recall, compared to 85% precision and 31% recall obtained by version 4.1 of Bitextor.

5.2 Cleaning Translation Memories Shared Task

Partner Prompsit participated in the NLP4TM shared task on cleaning translation memories.[10] Participants in this task were required to take pairs of source and target segments from translation memories and decide whether

[8] `http://www.statmt.org/wmt16/bilingual-task.html`

[9] https://sf.net/p/bitextor/wiki/

[10] `http://rgcl.wlv.ac.uk/nlp4tm2016/`

they were right translations. The task comprised three language pairs: English–Spanish, English–Italian and English–German, and three classification tasks for each pair:

- Binary I: decide whether the translation needs post-editing or not.
- Binary II: decide whether the translation corresponds to the source segment (even though it may need slight post-editing) or not.
- Fine grained. Choose among three categories: the translation does not need post-editing, the translation corresponds to the source segment but it requires slight post-editing, or the translation does not correspond to the source segment.

Our submission is based on a supervised classifier trained solely on the provided training data. The features we used can be split in two groups: those that represent the lexical similarity of the two halves of a parallel segment and make use of a probabilistic bilingual dictionary, and those that do not use linguistic information at all and are based on shallow properties such as sentence length, capitalized words, punctuation marks, etc.

Given a bilingual dictionary whose source language is *L1* and target language is *L2* and a pair of segments s_1, written in *L1*, and s_2, written in the *L2*, we computed the lexical similarity features described next. The feature *DICT-QMAX-L1* is defined as the product, for each word w in s_2, of the maximum translation probability from any word in s_1 to w according to the bilingual dictionary. The feature *DICT-QMAX-L2* is computed in the opposite direction (with the help of a bilingual dictionary whose source language is *L2* and target language is *L1*). We also used two additional features that account respectively for the proportion of words in s_1 and s_2 that can be found in the bilingual dictionaries.

Shallow features included: number of tokens in each segment, proportion between them, average token length (in characters) in each segment, number of punctuation marks in each segment, number of numerical expressions in each segment that can be found in the other segment of the pair, number of capitalized tokens in one segment that can be found in the other segment of the pair, etc.

We trained a support vector machine classifier with a radial-basis-functions kernel. We used the default parameters of the SVM implementation in the

Scikit-learn library.[11] The bilingual dictionaries (we built 6 bilingual dictionaries: one for each language pair and direction) were obtained from all the available parallel corpora at http://opus.lingfil.uu.se/. Each corpus was word-aligned by means of MGIZA++,[12] alignments were symmetrized, and the probabilities in the bilingual dictionaries were estimated by maximum likelihood from the symmetrized alignments.

As Prompsit's system was submitted after the task was closed, it was not included in the working notes and official results of the shared task.[13] However, shared task organisers sent us an updated version of the results which included our submission. The results were as follows:

- In English–Italian and English–Spanish, our system was either the best performing one or the second one, depending on the particular classification task (binary or fine-grained).
- In English–German, our system was ranked in position 5–6 (depending on the particular classification task) out of 7 participants.

6 Conclusion

This deliverable has covered the work done in the area of acquisition (work package WP3) during the period between the third (M36) and the fourth milestone (M48) of the project. Work performed include: the crawling of the Republic of Ireland top-level domain and the acquisition of additional parallel corpora, the development of methodology for the computer-assisted enrichment of rule-based machine translation dictionaries, the automatic inference of structural transfer rules for rule-based machine translation and their weighting to select the best rule in case of ambiguity, and the participation of Abu-Matran in two shared tasks: a document alignment shared task and a translation memory cleaning task.

References

Victoria Arranz, Khalid Choukri, Olivier Hamon, NÃºria Bel, and Prodromos Tsiavos. PANACEA project deliverable 2.4, annex 1: Issues related to

[11] http://scikit-learn.org/

[12] https://github.com/moses-smt/mgiza.git

[13] http://rgcl.wlv.ac.uk/nlp4tm2016/working-notes-on-cleaning-of-translation-memories-shared-ta

data crawling and licensing. http://cordis.europa.eu/docs/projects/cnect/4/248064/080/deliverables/001-PANACEAD24annex1.pdf, 2013.

Christian Buck and Philipp Koehn. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation*, pages 554–563, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W16/W16-2347.

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, Víctor M. Sánchez-Cartagena, Juan A. Pérez-Ortiz, Felipe Sánchez-Martínez, Mikel L. Forcada, and Rafael C. Carrasco. An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation Translation*, pages 19–26, Dubrovnik, Croatia, June 2014.

Miquel Esplà-Gomis, Nikola Ljubešić, Sergio Ortiz Rojas, Vassilis Papavassiliou, Prokopis Prokopidis, Víctor M. Sánchez-Cartagena, and Antonio Toral. Deliverable D3.1c acquisition for cycle 3. Technical report, Abu-MaTran, MSCA-IAPP project PIAP-GA-2012-324414, 2015.

Miquel Esplà-Gomis, Rafael C. Carrasco, Víctor M. Sánchez-Cartagena, Mikel L. Forcada, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. Assisting non-expert speakers of under-resourced languages in assigning stems and inflectional paradigms to new word entries of morphological dictionaries. *Language Resources and Evaluation*, pages 1–29, 2016. ISSN 1574-0218. doi: 10.1007/s10579-016-9360-9. URL http://dx.doi.org/10.1007/s10579-016-9360-9.

Miquel Esplà-Gomis, Mikel Forcada, Sergio Ortiz Rojas, and Jorge Ferrández-Tordera. Bitextor's participation in WMT'16: shared task on document alignment. In *Proceedings of the First Conference on Machine Translation*, pages 685–691, Berlin, Germany, August 2016. Association for

Computational Linguistics. URL http://www.aclweb.org/anthology/W/W16/W16-2367.

M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, 2011. Special Issue: Free/Open-Source Machine Translation.

Mikel Lorenzo Forcada, Miquel Esplà-Gomis, and Juan Antonio Pérez-Ortiz. Stand-off annotation of web content as a legally safer alternative to crawling for distribution. *Baltic Journal of Modern Computing*, 4(2):152–164, 2016.

Nikola Ljubešić and Filip Klubička. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

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, 2015.

Nikola Ljubešić, Filip Klubička, and Damir Boras. Inflectional lexicon hrLex 1.2, 2016a. URL http://hdl.handle.net/11356/1072. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Filip Klubička, and Damir Boras. Inflectional lexicon srLex 1.2, 2016b. URL http://hdl.handle.net/11356/1073. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. ISBN 978-2-9517408-9-1.

Vassilis Papavassiliou, Prokopis Prokopidis, and Stelios Piperidis. The ILSP/ARC submission to the WMT 2016 Bilingual Document Alignment

Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 733–739, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W16-2375.pdf.

Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. Parallel Global Voices: a Collection of Multilingual Corpora with Citizen Media Stories. 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). URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/778_Paper.pdf.

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. rulearn: an open-source toolkit for the automatic inference of shallow-transfer rules for machine translation. *The Prague Bulletin of Mathematical Linguistics*, 106:193–204, 2016. ISSN: 0032-6585.

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, and Juan Antonio Pérez-Ortiz. Source-language dictionaries help non-expert users to enlarge target-language dictionaries for machine translation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Prodromos Tsiavos, Stelios Piperidis, Maria Gavrilidou, Penny Labropoulou, and Tasos Patrikakos. Qtlaunchpad public deliverable d4.5.1: Legal framework. http://www.qt21.eu/launchpad/system/files/deliverables/QTLP-Deliverable-4_5_1_0.pdf, 2014.