Free/Open Source Shallow-Transfer Based Machine Translation for Spanish and Aragonese

Juan Pablo Martínez Cortés, Jim O'Regan, Francis M. Tyers

GTC, I3A, Universidad de Zaragoza
`jpmart@unizar.es`

Eolaistriu Technologies
`joregan@gmail.com`

DLSI, Universitat d'Alacant
`ftyers@dlsi.ua.es`

Abstract

This article describes the development of a bidirectional shallow-transfer based machine translation system for Spanish and Aragonese, based on the Apertium platform, reusing the resources provided by other translators built for the platform. The system, and the morphological analyser built for it, are both the first resources of their kind for Aragonese. The morphological analyser has coverage of over 80%, and is being reused to create a spelling checker for Aragonese. The translator is bidirectional: the Word Error Rate for Spanish to Aragonese is 16.83%, while Aragonese to Spanish is 11.61%.

Keywords: Aragonese, morphological analysis, machine translation

1. Introduction

Aragonese is a Romance language spoken in the northern area of Aragon (the southern flank of the central Pyrenees) by a population of about 10,000 speakers and an indeterminate number of second-language speakers. Although historically spoken in almost all Aragon, it has suffered a constant decline and progressive substitution by Spanish since the 15th century. In recent decades, the number of native speakers has dramatically decreased due to lack of intergenerational transmission. In most areas, only older people use the language. In contrast, there is a certain interest among young and mid-age people to learn Aragonese even in areas where the language is already lost as a native language. Recently, Aragonese has been legally recognised,[1] together with Catalan, with a status of "lenguas propias e históricas de Aragón",[2] and as one of the languages of application of the European Charter for Regional or Minority Languages in Spain. Because of the low number of speakers and its difficult sociolinguistic situation, it is one of the language with the lowest vitality in southwestern Europe, and one of those with fewer resources and opportunities for the future. UNESCO considers Aragonese a "definitely endangered" language (UNESCO, 2009).

Although the language is fragmented into several historical dialects, a loosely-defined and still evolving compositional standard[3] has been outlined in the last thirty years (Berceo, 2003; Metzeltin, 2004; Estudio de Filolochía Aragonesa, 2010). However, the absence of official normalisation and lack of language teaching favours the profusion of language models, which can hinder the identification of the speakers themselves with the different models (Paricio and Martínez, 2010). Even at the orthographic level, at least three spelling systems have been defined to write standard and dialectal Aragonese. This is, of course, a complication for developing tools for the language.

In the development of the Spanish-Aragonese language pair translator, we have selected the 2010 Orthographic Proposal of the "Academia de l'Aragonés" (Estudio de Filolochía Aragonesa, 2010). Though it has not reached full consensus, it is the most widely used standard in the generation of new on-line content in Aragonese (Wikipedia, `http://www.arredol.com`), and is predominant among active online users (blogs, twitter). Moreover, other linguistic tools, such as a spellchecker, are being developed using this spelling system. On the other hand, we have tried to be more inclusive when translating from Aragonese into Spanish: the aim is that the system is able to analyse as much dialectal morpho-lexical variation as possible, and, to a lesser extent, orthographic variations.

The rest of this article is outlined as follows: first, we describe the development of the system, and the resources used and created; we then describe the status of the system, describing the coverage of the morphological analyser, and an evaluation of the translator for the purpose of post-editing; finally, we describe future work for the improvement of the system.

2. Development

The system is based on the Apertium platform (Forcada et al., 2011); a Free/Open Source platform for shallow transfer machine translation. The platform was originally designed for the Romance languages of Spain, so no deviation from the usual design of an Apertium-based translator was required. As well as the platform, the linguistic data of the translators are also available under the terms of the GNU General Public License.

Development to date has consisted of two phases: the first for assimilation, from Aragonese to Spanish, to be used as a tool in preparing the second phase; the second phase, to extend coverage, to make the system bidirectional, adding a module for orthographical operations (such as contractions), and to ensure adherence to the orthographical standard. Development took place in short bursts over the course of two years: 3 weeks for the first phase, 7 weeks for the second. Both versions of the system are available for download at the Apertium development site.[4]

[1] Under the so-called "Act on Aragon Languages", speakers achieved a minimal legal recognition. However, the Act, which established a language regulating body (Academy) and voluntary classes at all educative levels in the regions where the language is still natively spoken, has hardly been developed.

[2] "Native and historical languages of Aragon"

[3] A standard language, composed from all of the dialects

[4] `http://sourceforge.net/projects/`

2.1. Resources

We were able to reuse several resources provided by various Apertium translators in the creation of this package. The Spanish monolingual data, and most of the transfer rules, were taken from the Spanish-Catalan package with minimal changes.

For Aragonese, there were few resources. The English edition of Wiktionary provided conjugations for some model verbs, from which we were able to build initial inflection paradigms. We used the Aragonese Wikipedia as a source of frequency lists, to guide the development of the morphological analyser.

During the initial phase of development, the Aragonese Wikipedia was in the process of switching to the new orthographical standard. Articles conforming to the standard were added to their own category, which we were able to use, via Mediawiki's export function, to create a subcorpus of standardised words. The standard orthography is now the norm on Wikipedia, and this category has been superceded by categories for articles written in each of nine dialects.

In lieu of a parallel corpus,[5] we used comparable material obtained by manually extracting the first sentences from a number of articles which had versions in both the Aragonese and Spanish editions of Wikipedia. In many cases, these were close enough to function as parallel sentences (see table 1); in others, they contained at least phrases that were useful for testing purposes.

Manual extraction of these sentences was sufficient for our needs, but for future development we wished to automate the process. Rather than duplicate existing work, we chose to reuse data made available by DBpedia(Auer et al., 2007). Although extracting linguistic data is not the primary aim of DBpedia, the "short abstract" datasets it provides contain the first paragraph of each article. We loaded the datasets for Spanish and Aragonese[6] into a database, and used a simple query (Table 3) to extract a comparable corpus consisting of the first paragraphs of articles which have an interwiki link to each other. The resulting dataset contains over 16,000 comparable abstracts. [7]. An example of a pair of comparable abstracts can be viewed in Table 2.

Table 1: Example first sentences from the Wikipedia articles "Mar".

es	Un mar es una masa de agua salada de tamaño inferior al océano
an	A mar u o mar ye una masa d'augua salada de grandaria inferior a l'ocián
	A sea is a body of salt water smaller than an ocean

Table 2: Example extract from the DBpedia abstract dataset.

es	John Joseph Nicholson es un actor, productor, guionista y director de cine estadounidense doce veces nominado y tres veces ganador del Premio de la Academia. En activo como actor desde 1958.
an	Jack Nicholson (nombre artistico de John Joseph Nicholson) ye un actor y director cinematografico estatounitense, naixito o 22 d'abril de 1937 en Nueva York.
Gloss es	*John Joseph Nicholson is an actor, producer, screenwriter and director of American cinema twelve times nominated and three times winner of the Academy Award. Active as an actor since 1958.*
Gloss an	*Jack Nicholson (artistic name of John Joseph Nicholson) is an American actor and director, born 22 April 1937 in New York.*

2.2. Dictionaries

The Aragonese morphological dictionary and the Spanish-Aragonese bilingual dictionary were created in a synchronous manner. Closed categories (prepositions, determiners, numerals) were manually added first, along with a few examples of cognates from the open categories. These cognates were used to build a set of common transformations, both for normalisation of non-standard forms, and for cognate induction; and to build a set of equivalent suffixes, which, as well as functioning as transformations, also served as a means of filtering words for equivalence, and for assigning categories: for example, *-dá, -dat, daz* and *-datz* all refer to the same feminine noun (*-dat, -datz* in the standard orthography), which typically have cognates with the suffix *-dad, -dades* in Spanish: for example, (*uniformidat, uniformidatz* ("uniformity", "uniformities") in Aragonese, *uniformidad, uniformidades* in Spanish.

Cognate induction was then performed by filtering words by suffix, applying the extracted transformations, and com-

Table 3: SPARQL query used to extract abstracts.

```
PREFIX rdfs:
<http://www.w3.org/2000/01/rdf-schema#>
SELECT ?es, ?an WHERE {
  ?subject rdfs:comment ?es.
  ?subject rdfs:comment ?an.
  FILTER (lang(?es) = "es"
       && lang(?an) = "an")
}
```

apertium/files/apertium-es-an/

[5] The text of "Ley d'Uso, Protección y Promoción d'as Luengas d'Aragón" became available towards the end of the first phase, though too late to be used. We chose not to use it in the second phase, to have genuine parallel text to use in evaluation.

[6] http://downloads.dbpedia.org/3.7/es/short_abstracts_es.nt.bz2 and http://downloads.dbpedia.org/3.7/an/short_abstracts_an.nt.bz2 for Spanish and Aragonese, respectively.

[7] Available from https://github.com/jimregan/es-an-sentences

Table 4: Comparison of dictionaries.

	0.1	0.2
Lemmas (bil)	8598	23597
Lemmas (mon)	8773	22748
Paradigms	119	517

paring the result with a filtered list of lemmas extracted from the Spanish analyser. In an additional step, the process was repeated to find a normalisation candidate, using the subcorpus of standardised Aragonese, where possible, or the most frequent form otherwise.

In comparing the words contained in the first version of the dictionary with the corrected version, we find a difference of 14%, split into 7.4% normalisation errors (non-standard version selected), and 6.6% outright errors (including misspelled words, non-Aragonese words, and incorrect Part of Speech). A further 0.6% of the remainder (0.52% of total) were due to cognate induction errors (false friends).

3. Status

3.1. Coverage

Lexical coverage is an important factor in Machine Translation, because of the effect that unknown words can have in the translation produced: as well as the word itself potentially not being translated correctly, the lack of concordance caused by not knowing the features of that word (such as gender and number) can lead to generation errors in the words which are required to agree with it grammatically.

We calculated the naïve coverage (the proportion of words which received at least one analysis) of the analyser, using Aragonese Wikipedia[8] as a corpus. Naïve coverage of the analyser on Wikipedia is 87.67% (over 3,648,449 words). In a corpus consisting of three short novels, naïve coverage is 92.15% (over 138,355 words).

3.2. Evaluation

To evaluate the translator, we used the text of "Ley d'Uso, Protección y Promoción d'as Luengas d'Aragón":[9] 5389 words of Aragonese, 5884 of Spanish.

As our aim is to produce a system suitable for post-editing, we chose to evaluate using *word error rate* (WER). WER is calculated based on the number of changes required to change the machine translated output into the reference text, which is a good indicator of the amount of work that will be required in post-editing.

In Table 5 we provide two figures for WER, both without unknown word marks ("WER"), and with ("WER$_2$"), because of the high number of unknown words identical in source and target (which do not require editing themselves, but may potentially lead to agreement or structural errors elsewhere).

The disparity between the es–an and an–es results can, in large part, be attributed to the looseness of Aragonese standard and, specifically, to different choices of definite articles. The standard permits some variation, such as "o" / "lo" or "a'l" / "a o", which is counted as an error of two words.

In addition to WER, we have also evaluated the translator using BLEU (Papineni et al., 2002). Although BLEU has been shown to favour SMT-based systems over rule-based systems (Callison-Burch et al., 2006), it is something of a de facto standard, and is expected in evaluations. We present the results in Table 6. As with WER, we present results both without unknown word marks ("BLEU"), and with ("BLEU$_2$").

4. Future Work

It is intended to continue to develop and maintain the system, to both extend the coverage of the lexicons and to improve the quality of the translations.

The Apertium Nynorsk and Bokmål system (Unhammer and Trosterud, 2009) utilised feedback from Wikipedia[10] in developing the system. We have received positive feedback from editors of the Aragonese Wikipedia, which we hope will lead to the availability of post-edited translations we can use to extend the system. To that end, we have begun work on an improved version of the "stupid aligner"[11] used to develop the Apertium Czech to Polish system (Ruth and O'Regan, 2011).

The morphological analyser contains some support for dialectal forms. We hope to extend this, and to enable generation of these forms. Independent of this, we feel that a translation-less mode of the system could be useful as a tool for standardising Aragonese text.

The list of standardised forms from the analyser has been contributed to an effort to create a spell-checking dictionary for Aragonese[12]. We hope that feedback obtained from this collaboration will contribute towards the extension of the system.

The corpus of sentences from DBpedia requires some further refinement, to be more useful for machine translation. Work has begun on the first stage of this, to create a sentence splitter for Aragonese.[13] Additionally, work has begun on localising the DBpedia extraction software to Aragonese, and the Aragonese Wikipedia. The current abstract extraction relies on interwiki links with the English Wikipedia: we feel that more data could be extracted by comparing direct interwiki links.

As can be seen in Table 2, the abstracts are not always as similar as the sentences we manually selected (Table 1).

[8] Specifically, `http://download.wikimedia.org/anwiki/20111007/anwiki-20111007-pages-articles.xml.bz2`

[9] `http://www.academiadelaragones.org/biblio/LEI\%20DE\%20LUENGAS\%20D'ARAGON.pdf`

[10] The Nynorsk Wikipedia has a category for articles derived from Apertium translations: `http://nn.wikipedia.org/wiki/Kategori:Omsett_med_Apertium`

[11] Available from Apertium SVN: `https://apertium.svn.sourceforge.net/svnroot/apertium/incubator/stupid-unknown-extractor`

[12] `https://addons.mozilla.org/es-ES/firefox/addon/corrector-ortografico-aragones/`

[13] As Apertium operates at phrase level, sentence-level tokenisation was not required by the translator.

Direction	WER	WER_2	Unknowns	Free Rides
es–an	16.83 %	19.37 %	4.78 %	53.10 %
an–es	11.61 %	14.12 %	5.67 %	44.05 %

Table 5: Evaluation results for both directions. Free rides are those unknown words which are identical in both the source and target language. Although they do not cause a degradation in translation quality, it is relevant to take them into account when evaluating the system. Unknown words are included as an indication of naïve coverage over the test sets. For brevity, we refer to the languages by their ISO 639-1 codes: `es` for Spanish, and `an` for Aragonese.

Direction	BLEU	BLEU_2
es–an	0.7149 %	0.6493 %
an–es	0.7863 %	0.7184 %

Table 6: BLEU scores for both directions.

SMT techniques have been applied to the task of to finding translated passages in comparable documents (Sánchez-Martínez and Carrasco, 2011). Lacking sufficient text to train an SMT system, we intend to "relearn" our RBMT system (Dugast et al., 2008) to investigate the possibility of using the search for translated passages among the abstracts as a means of removing less parallel abstracts.

Work has also begun on adapting DBpedia Spotlight (Mendes et al., 2011) to both Spanish and Aragonese. DBpedia Spotlight is a semantic annotation tool which adds semantic annotations from DBpedia to unannotated text. To achieve this, the indexing portion of the software first extracts a set of lexicalisations for each topic from Wikipedia. First, the article titles themselves; second, the text used in each link within Wikipedia. The index itself links each article to a paragraph of context surrounding the link, which is used to disambiguate links by treating disambiguation as a document search problem.

Both the lexicalisation data and the index represent interesting sources of linguistic data. The lexicalisation data from a pair of Wikipedias, when combined with the interwiki links from DBpedia, provides us with clusters of synonyms that can be queried in a database. We have had promising initial results in extracting single word cognates from these clusters, but we are interested in using it as a potential source of multiwords. The index itself represents a source of data for lexical selection rules: by mapping ambiguous entries in the lexicon to their Wikipedia article title, we hope to extract local context with which to infer these rules.

5. Acknowledgements

The authors wish to thank Mikel Forcada for his feedback on early drafts of this paper; we also wish to thank the anonymous reviewers for their useful comments.

The work on extracting comparable sentences from DBpedia was performed during a project to create a Basque to English machine translation system, for which we acknowledge the funding of the European Association for Machine Translation.

6. References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, chapter 52, pages 722–735. Springer Berlin / Heidelberg, Berlin, Heidelberg.

R. Berceo. 2003. Normativisation, a priority for Aragonese. `http://www.aber.ac.uk/mercator/images/Bercero.pdf`.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the Role of BLEU in Machine Translation Research. In *Proceedings of EACL 2006, 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2008. Can we relearn an RBMT system? In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 175–178, Stroudsburg, PA, USA. Association for Computational Linguistics.

Estudio de Filolochía Aragonesa. 2010. Propuesta ortográfica de l'academia del aragonés. `http://www.academiadelaragones.org/biblio/EDACAR7_2.pdf`.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez Felipe Sánchez-Martínez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*. OnlineFirst.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding Light on the Web of Documents. In *7th International Conference on Semantic Systems (I-Semantics)*.

M. Metzeltin. 2004. Las lenguas románicas estándar. historia de su formación y de su uso.

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.

Santiago Paricio and Juan Pablo Martínez. 2010. New ways of revitalization for minority languages: the impact of the internet in the case of Aragonese. `http://www.uoc.edu/ojs/index.php/digithum/article/download/n12-paricio-martinez/n12-paricio-martinez-eng`.

Joanna Ruth and Jimmy O'Regan. 2011. Shallow-transfer rule-based machine translation for Czech to Polish. In *Second International Workshop on Free/Open-SourceRule-Based Machine Translation*, pages 69–76, Barcelona.

Felipe Sánchez-Martínez and Rafael C. Carrasco. 2011. Document translation retrieval based on statistical machine translation techniques. *Applied Artificial Intelligence*, 25(5):329–340, May.

UNESCO. 2009. Interactive world atlas of endangered languages. `http://www.unesco.org/culture/ich/index.php?pg=00206`.

Kevin Unhammer and Trond Trosterud. 2009. Reuse of Free Resources in Machine Translation between Nynorsk and Bokmål. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 35–42, Alicante.