Language Technology for Language Communities: An Overview based on Our Experience

Iñaki Alegria, Kepa Sarasola
IXA group, University of the Basque Country (UPV/EHU)
Informaika Fakultatea, Lardizabal 1, 20018 Donostia. Basque Country
[ì.alegria@ehu.eus]

Abstract

IXA is a research group that has been working on language technology, mainly on Basque, during the last 28 year. As a result of years of collaboration with the Basque community and communities related to other languages we conclude that Language Technology to be an important factor for language development, previously (or in parallel) an initial core work is needed: 1) standardization and 2) generation of open contents. Bearing in mind these requisites, we propose the definition of a BLARK (Basic Language Resource Kit) to identify a minimal set of basic resources, and then we suggest tools for their adaptation to different languages depending on the size of their speakers' community and digital resources.

Introduction

Ixa group (www.ixa.eus) is a research group created in 1988 with the aim of laying foundations for research and development of Natural Language Text-Processing (NLP) and Human Language Technology (HLT) for Basque language. Now it is a big multidisciplinary group composed of computer scientists and linguists.

Two distinguishing features of the Ixa Group are that it deals with a less resourced language (Basque) and that it combines classic linguistic modelling and data analysis with innovative probabilistic and machine learning approaches to NLP.

At the very beginning, thirty years ago, our first funding was associated to the creation of a translation system for Spanish-Basque. But after some preliminary studies we realized that it was more important to concentrate our efforts in creating basic tools and resources for Basque (morphological analyser/generator, electronic dictionaries, annotated corpora, semantic databases...) that later on could be used to build many other general language applications, rather than creating an *ad hoc* and extremely complicated MT system. This thought was the seed to design our strategy to make progress in the adaptation of Basque to Language Technology.

Nowadays our research has resulted in state-of-the-art technology for robust, broad-coverage natural-language processing for Basque. These technologies/resources include a spelling checker (Xuxen), Basque Wordnet (BasqueWN), the corpus of Science and Technology (ZT corpus), a syntactically annotated corpus (EPEC), a Spanish-Basque MT system (Matxin), a NLP pipeline for text processing (Ixa-pipes) and an opinion-mining tool (Behagunea).

Based on our experience on NLP for less-resouced languages (Alegria et al., 2011), we have been collaborating for many years with two kinds of language communities:

- The community working in the socialization of the Basque language (dictionaries, language learning methods, Wikipedia, keyboards and interpretation tools for smart-phones...)
- Other linguistic communities with less resource, in order to help them in the technological development of their language (Quichua, Nahuatl, Spanish in Cuba...).

Borin (2009) pointed to the promise of the HLT for lesser-known languages and describes the linguistic diversity in the information society. He cites the paper from Ostler "*a language will not get by in the world of today unless it is equipped with a parser and a multi-million-word corpus of text*". He analysed the relation among the sociology of language and HLT, and gave us some strategic considerations.

In our opinion technology may be an important factor for language development, but there is a core work which had to be implemented before (or in parallel):

- Standardization: the fragmentation of the community in dialects makes it difficult the generation of written contents. Standardization has to be a priority in the way to effectively promote the use and to give prestige to the language. Dialects, of course, have their role, specially in oral contents and informal uses.
- Digital contents: without a minimum basis (mainly scholar books, translations and Wikipedia) it will be impossible the generation of interesting tools for the language community.
- Open contents and open source software: the decision to promote the production of open contents and the use open source tools is a capital strategy in order to ensure an incremental and sustainable development of this technology.

Bearing in mind these requisites we present the concept of BLARK and its adaptation to the size of the community. A BLARK for a language (Krauwer, 2003) is the minimal set of basic resources (software modules, corpora, dictionaries, etc.) that is necessary to do further research and development in the field of Language Technology.

The paper is structured as follows. After discussing the relevance of several elements cuch as the role of a language community, the level of standardization and the amount of text evailable (Section 2), we present related work (Section 3) In Section 4 we present the key resources and applications to be implemented in a concrete roadmap for low-resourced languages, including corpus compilation, digital dictionary, spelling checker, morphology, corpus annotation, POS tagger and text-mining. Finally, in Section 5, we draw conclusions.

Relevance of community, standardization and digital content

The standardization of the language is a previous requisite for a successful use of the written language.[1]

In Basque there are approximately 800,000 speakers and six dialects. The dialects are very distinct from each other. In 1968 the Basque Academy of the Language decided to create the Standard Basque. After some years of discussions, finally it was widely accepted and now it is the standard Basque (named '*Batua*') the language model used in (allmost) all the formal texts: school, university, administration, official pages in Internet... TV and radio journalists and academics speak in a standard way.

As Hualde and Zuazo (2007) say "*By any criterion that we may choose, the standardization of Basque in recent years has been a very successful project. Nowadays, standard Basque, which was not developed until the late 1960s, is used in education at all levels, from elementary school to the university, on television and radio, and in the vast majority of all written production in Basque. This success in the societal acceptance of standard Basque is most remarkable given the fact that there is no administration common to all territories where Basque is spoken (divided as they are between Spain and France and even, within Spain, into two separate administrative regions with different legislation regarding the Basque language) and that Basque speakers are almost always fully bilingual in either Spanish or French, so that the existence of a standard Basque language is not strictly required for communication beyond the local level.*"

We want to underline the relevance of the work done by the linguistic community in this process; it was the community who pressed for an academic/political decision to accept the standard, and it was the community who generated new resources using the standard (books, magazines, dictionaries, a newspaper, wikipedia…).[2] It has been specially important the role of the Basque schools (*Ikastolak*) in the recovery and standardization of Basque (Lopez-Goñi, 2003).

It was very important for us the fact that the standard Basque had been defined and widely accepted before our research group started to develop new NLP tools or applications. When we needed linguistic knowledge we did not need to create it by ourselves, this work had been done previously. We had no need to deal with different dialectical variants for a word, no need to choose one of those variants, the Academy of Basque (Euskaltzaindia) had done it before.

Later on when we have been collaborating with academics or other communities in order to develop technology for low-resourced languages we were more aware of the importance of the standardization of a language. For example joining forces for Quechua is a difficult task because in Peru, Bolivia and Ecuador they use different variants of the same language.[3]

In addition, nowdays, when we need corpora for learning or for inference, it is easer finding adequate text because of the increment of written production in Internet. Consequently this aspect has become a key factor for success because text-corpus is the raw material for the present main technological paradigm: data-driven language engineering.

As we will explain below Wikipedia is becoming a key resource, not just as a single text corpus but even as a suitable basis for the development of new tools and applications. Unfortunately, sometimes there is not a common agreement between local communities for defining and promoting a standard variant of the language. The consequence uses to be a smaller wikipedia, inefficient diversification of human resources, and a more divided community, i.e. using classical Nahuatl or not is still an open discussion.[4]

Dialects and variants are also an important matter for the language community,[5] but in our opinion standard language is a priority for text processing.

BLARK and open source

Krauwer (2003) proposed a "Basic LAnguage Resource Kit (BLARK)" as a roadmap of tools to be developed for each language using the terminology defined in a joint initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association) in 1998. In all these works a list of basic resources and tools are listed. The term BLARK has been very successful and it is used in a large number of papers in the area.

Streiter et al. (2006) report on HLT projects for non-central languages and proposes instructions for funding

[1] It may be argued that it is not a need for speech processing but most of the speech-to-text systems need resources based on standard texts.

[2] *Garabide* (http://www.garabide.eus/english) it is a NGDO which try to help language communities using the revitalization of the Basque language as a model for them.

[3] We know that the variants can be considered as different languages, and here communities have to decide if they prefer join efforts or work separately.

[4] https://meta.wikimedia.org/wiki/Proposals_for_closing_projects/Closure_of_Classical_Nahuatl_Wikipedia

[5] Social networks (specially Twitter) is becoming also an important resource for identification/treatment of variants/dialects

bodies and strategies for developers. They use the non-central term and underline the importance of making use of free software to improve the results. The chapter about benefits and unsolved problems when using open source software for non-central languages is very interesting. Forcada (2006) remarks the opportunity of using open source machine translation for minor languages.

The ELSNET network of excellence prepared definitions for a language resources and evaluation roadmap. The elements in the diagram (HLT products) are classified into three subsets: Language Resources, Language Processing, Language Usage. (Language Resources, Language Tools and Language Applications in our proposal).

Based on several indicators we have proposed six levels in order to classify rhe adaptation of the languages to the technology (Alegria et la., 2011):

1. English: Around 45% of the web pages are written in English. Almost all the HLT applications are available for English. The most of the research is carried out testing on English texts.
2. Other top 10 languages that cover almost 50% of Internet users. There are the languages for which active resource development continues and the most major companies on Internet support them. Streiter et al. (2006) call the central-languages.
3. Around 70 languages with any HLT resources registered. Sometimes they are named non-central languages.
4. Around 300 languages with any lexical resource on-line registered in yourdictionary.com. It is almost the same set of the languages that are in Wikipedia or the set of languages that have defined their standard. The term low-resourced (or lesser-resourced) language is used to be applied to these languages (and to the previous level also).
5. Around 2,000 languages that have writing systems (Borin, 2009).
6. The big bag also including only-spoken languages in the world (more than 4,000). The most of them can be considered endangered languages.

In the next section we try to fix, according to our experience, the most important resources tools and applications to be developed as a roadmap for the languages in the range 4-5.

In addition to this we want to stress how the linguistic and academic communities can cooperate in their development. In some cases, i.e. natural disasters, it can be interesting a quick response (Munro, 2010), but in general is better to set a plan depending on the situation of the language: number of speakers, connectivity of them, digital resources, integration in the school...

If there is an important group of Internet users collaborative tools are a very productive way. Tools on Wikimedia (Wiktionary, Wikipedia...) are the most known, but there are other tools as the crowdsourcing platforms (Sabou et al., 2012) which can be used by language communities. When Internet users are scarce finding collaboration from academia and schools is more suitable

Key resources and applications

In the next subsections we propose a concrete roadmap for low-resourced languages, beginning from the most basic resources/tool/applications. We have selected mainly open-source resources and tools. This roadmap is based in our experience and the proposal by Streiter et al., (2006).

We will not include machine translation among the applications because it need more resources than those that have languages in the range 4-5. Anyway if there are close languages with more resources a machine translator among similar languages can be built without big effort. Apertium[6] (Forcada, 2006) is a nice example in this area.

Corpus compilation and digital dictionary

Corpus. A monolingual corpus is the first basic resource for language technology. Its most important feature is the size but there are other features to be taken into account: normalization/variants, domain, single/multiple sources... It can be a big project if we want to build a "national corpus" or a "monitor corpus" including metadata (XML/TEI is the standard way for this) and additional tools.

Wikipedia is a nice option for corpus extraction, but in the cases where wikipedia does not exist for a language or it is too short, dealing with *web as a corpus* techniques may be a good option if substantial texts are available in Internet. In other case, scanning texts or collaboration with editors and teachers/academics are the remaining option.

Web as a corpus techniques were described by Kilgarriff and Grefenstette (2003), and Webcorp[7] is a interesting tool for this aim. Sometimes some adaptations of the program to the particular linguistic features of the language should be performed (Leturia et al., 2007).

When scanning of documents or compilation of digital files are necessary, it is important to preview and measure the real dimension of the work: compiling documents or files in different formats, dealing with licences and legal issues, scan or format conversion, OCR, insertion of metadata... From our experience (Areta et al., 2007) this is a big work, much bigger than what was previously expected. Gutierrez-Vasques et al. (2016) show an example of a bilingual compilation.

There are also global projects for building corpus for multiple languages (Abney and Bird, 2010, Scannell, 2007).

Based on the corpus first application can be developed, for instance, examples for language learning, dictionary of frequencies, basic games (looking for short words,

[6] https://www.apertium.org

[7] http://www.webcorp.org.uk/live/

long words, palindromes...). Natural Language Toolkit (NLTK)[8] is a very interesting tool set for the development of such applications.

Digital dictionary. A dictionary is a key tool for students. A very important tool. From our experience, together with the spelling corrector, it is the most practical application that we have developed. When available Wiktionary can be the basis, but it can also be built from a corpus or from a previous dictionary[9].

Corpus may be helpful for quality testing and to find new entries, but the best option is a previous lexicographic work. From our experience we know that in some communities a digital dictionary exists, but it is not available from Internet or it has a proprietary licence. A very important task is the conversion of this dictionaries into a multimedia online dictionary based in a lexical database.

A good experience for us was the semiautomatic transformation of the Cuban "Diccionario Básico Escolar" (Miyares et al., 2010)[10].

For Basque, Euskalbar[11] (an add-on for Internet navigators which send concurrent queries to existing online dictionaries and corpora, and show all the results simultaneously) is a key application for the community.

Based on the dictionary, new applications can be developed, specially for students. In that way, our group was involved in building the Basque version of *Apalabrados*[12].

Spelling and Morphology

As we said before the spelling checker is one of the most successful applications for a language. Students, teachers, journalists, writers... use to use it. It is even more necessary when the written-system for a language is in development. Furthermore, in the case of Basque it has been a very effective tool in the standardization process.

A spelling checker may be generated from a big (good) corpus, but its quality and coherence would be better if its construction were based on a morphological analyser. It is mandatory for morphologically-rich languages.

A morphological analyser obtains, for each word, its possible morphological segmentations, mainly lemma and part-of-speech category associated to each word-form. Based on it the speller decides that words without morphological analysis are mistakes or variants.

To build the analyser it is necessary to specify: (1) the set of lemmas with their categories, (2) the affixes, (3) the morphotactics describing valid linkings among lemmas and affixes and (4) the morphophonological changes produced when linking lemmas and affixes. The first specification, the set of lemmas, may be obtained from the digital dictionary and the others from academics or from a formal basic grammar. For putting all together there are some tools; we used the two most popular tools: foma[13] and hunspell.[14] The first one (Hulden, 2009) is linguistically better motivated and simpler for the description, but using the second one has been more successful because the description can be directly integrated as a speller in a lot of software packages (Libreoffice, Mozilla…).[15] For Basque (Alegria et al., 2009) and for Quichua (Rios, 2011) both options have been combined, by creating the first description using foma and then automatically converting it to hunspell.[16]

Of course, the community has an important role to play in the construction and distribution of he Spelling checker: testing the tool, spreading it, and helping new users to install in their computers, sending feedback on errors or missing lemmas....

Annotation, POS tagging and text-mining

Raw text corpora are a nice resource to develop very basic NLP applications, but corpora annotated with morphological, syntactic or word meaning information opens the door to (semi-)automatically build part-of-speech taggers, and tools for text mining.

For instance, we built EPEC (Reference Corpus for the Processing of Basque) for Basque[17], which is a 300,000 word corpus of standard written Basque It was manually tagged at different levels: morphosyntax, syntactic phrases... It has already been used for the construction of some tools such as a POS tagger

The POS tagger is another key tool together with the digital dictionary and the spelling checker, because it is a mandatory previous step for text mining: fact extraction, identification of entities (persons, places, organizations), extraction of terminology, text simplification... The tagger assigns to each word in a text its part-of-the-speech, based on its definition (or morphological analysis) and its context.

As we have said before based on POS tagging it is possible to build a lot of applications for text mining, but more powerful tools too. IXA pipes (Agerri et al., 2014) framework is an example of how built easily these new tools. It is a modular set of Natural Language Processing tools (or pipes) which provide easy access to NLP technology for several languages. It offers robust and efficient linguistic annotation very useful in text-mining. This open technology is easily adaptable to any other language, the only requisite is the access to linguistically annotated corpus.

[8] http://www.nltk.org/

[9] yourdictionary.com presents links to on-line lexical resources (http://www.yourdictionary.com/languages.html for 307 languages

[10] http://ixa2.si.ehu.es/dbe

[11] https://addons.mozilla.org/eu/firefox/addon/euskalbar/

[12] http://www.apalabrados.com/

[13] https://fomafst.github.io/

[14] h.ttp://hunspell.github.io/

[15] https://addons.mozilla.org/en-US/firefox/language-tools/ List of the spelling-checkers supported by Mozilla

[16] Another matter is that Microsoft Office is the main tool for a lot of users. Streiter et al. (2006) discuss it.

[17] Our steps on standardization of resources brought us to adopt TEI and XML standards as a basis for linguistic annotation (Artola et al., 2009).

Conclusions

Language technology is a powerful help for the communities related to low-resourced languages, in order to revitalize the language and to effectively promote the use of their language.

But there are some prerequisites to allow Language technology to be used. A language community that will activate the distribution and dissemination of the LT tools is needed The existence of a standard for the language and a wide acceptance of it will definitively make easier the development of new NLP tools and their effectiveness.

Corpus compilation, digital dictionary, spelling checker, morphology, corpus annotation, POS tagger and text-mining are the first steps to be faced. We have presented our fruitful experience dealing with Basque, and some suggestions for other languages that want to design a roadmap for language technology.

Acknowledgements

The authors of this work has received financial support from theses two projects: TADEEP (MINECO-Spain, TIN2015-70214-P) and ELKAROLA (Basque Country, KK-2015/00098) .

References

Abney, S., Bird, S. (2010). The human language project: building a Universal Corpus of the world's languages. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 88-97). Association for Computational Linguistics.

Agerri, R,. Bermudez,J., and Rigau, G. (2014): "IXA pipeline: Efficient and Ready to Use Multilingual NLP tools", in: Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), 26-31 May, 2014, Reykjavik, Iceland.

Alegria, I., Artola, X., De Ilarraza, A. D., & Sarasola, K. (2011). Strategies to develop Language Technologies for Less-Resourced Languages based on the case of Basque. Proceedings of 5th Language & Technology Conference: HLT as a Challenge for Computer Science and Linguistics. pp: 42-46, November 24-27, 2011, Poznan.

Alegria, I., Aranzabe, M., Arregi, X., Artola, X., Díaz de Ilarraza, A., Mayor, A. and Sarasola, K. (2011). Valuable Language Resources and Applications Supporting the Use of Basque. Z. Vetulani (Ed.): LTC 2009, Lecture Notes in Artifitial Intelligence LNAI 6562, pp. 327--338. Springer, Heidelberg.

Alegria, I., Etxeberria, I., Hulden, M., Maritxalar, M. (2009). Porting Basque morphological grammars to foma, an open-source tool. In International Workshop on Finite-State Methods and Natural Language Processing (pp. 105-113). Springer, Berlin.

Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Diaz de Ilarraza, A., Ezeiza, N., Sologaistoa, A. (2007). ZT Corpus: Annotation and tools for Basque corpora. Proceedings of Corpus Linguistics 2007.

Borin, L.(2009). Linguistic diversity in the information society. SALTMIL2009 Workshop: IR-IE-LRL. Information Retrieval and Information Extraction for Less Resourced Languages. University of theBasque Country.

Forcada, M. (2006). Open source machine translation: an opportunity for minor languages. In Proc. of the Workshop Strategies for developing machine translation for minority languages, LREC (Vol. 6, pp. 1-6).

Gutierrez-Vasques, X., Sierra, G., Pompa, I. H. (2016). Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl. In LREC 2016.

Hualde, J. I., Zuazo, K. (2007). The standardization of the Basque language. Language Problems and Language Planning, 31, 143-168.

Hulden, M. (2009). Foma: a finite-state compiler and library. In Proceedings of the 12th Conference of the EACL: Demonstrations Session (pp. 29-32). Association for Computational Linguistics.

Kilgarriff, A., Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. Computational linguistics, 29(3), 333-347.

Krauwer, Steven. "The basic language resource kit (BLARK) as the first milestone for the language resources roadmap." Proceedings of SPECOM 2003 (2003): 8-15.

Leturia, I., Gurrutxaga, A., Alegria, I., and Ezeiza, A. (2007). CorpEus, a 'web as corpus' tool designed for the agglutinative nature of Basque. In Building and exploring web corpora, Proceedings of the 3rd Web as Corpus workshop (pp. 69-81).

López-Goñi, I. (2003). Ikastola in the twentieth century: an alternative for schooling in the Basque Country. History of Education, 32(6), 661-676.

Miyares E., Ruiz-Miyares l., Álamo C. Pérez C, Artola X., Alegría I. Arregi X. La segunda y tercera ediciones del Diccionario Básico Escolar. Proceedings of the 14th EURALEX International Congress; 519-526.

Munro, R. (2010). Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. AMTA Workshop on Collaborative Crowdsourcing for Translation (pp. 1-4).

Rios, A. (2011). Spell checking an agglutinative language: Quechua. University of Zurich. Zurich Open Repository and Archive.

Sabou, M., Bontcheva, K., Scharl, A. (2012, September). Crowdsourcing research opportunities: lessons from natural language processing. In Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies (p. 17). ACM.

Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop (Vol. 4, pp. 5-15).

Streiter, O., Scannell, K. P., and Stuflesser, M. (2006). Implementing NLP projects for noncentral languages: instructions for funding bodies, strategies for developers. Machine Translation, 20(4), 267-289.