Opentrad: bringing to the market open-source based Machine Translators

Ibon Aizpurua Ugarte [1], *Gema Ramírez Sánchez* [2], *Jose Ramom Pichel* [3], *Josu Waliño* [4]

[1] Eleka Ingeniaritza Linguistikoa, S.L.
[2] Prompsit Language Engineering
[3] Imaxin | Software
[4] Elhuyar Fundazioa
ibon@eleka.net, gema@prompsit.com, jramompichel@imaxin.com, josu@elhuyar.com,

Abstract

Most successful machine translation (MT) systems built until now use proprietary software and data, and are either distributed as commercial products or are accessible on the net with some restrictions. This kind of MT systems are regarded by most professional translators and researchers as closed and static products which cannot be adapted or enhanced for a particular purpose. In contrast to these systems, we present Opentrad, an open-source transfer-based MT system intended for related-language pairs and not so similar pairs. The project is funded by the Spanish government and shared among different universities and small companies. It uses different translation methods according to each language pair. For related-languages it uses shallow-transfer, even though for non-related pairs the system uses deep-transfer. The translation speed obtained is very high because it uses a finite-state transducer technique. The novelty of Opentrad consists of an introduction of open source software-development methodology and interoperability of standards in the field of MT.

Index Terms: machine translation, open source, business

1. Introduction

Most successful MT systems built until now use proprietary software and data, and are either distributed as commercial products or accessible on the net with some usage restrictions. Most professional translators and researchers view them as closed products since they cannot be easily adapted to particular purposes, integrated into other applications or used as resources in research or development projects. Besides, these MT systems mostly use *ad hoc* formats for linguistic data which are unreadable and very hard to maintain or extend.

All these aspects of commercial systems have a negative impact on the development of new techniques or the addition of new language pairs. A better concurrence between developers would had led to a positive motivation to improve existing MT systems, but making new systems from scratch is so costly that usually the primary goals are constricted by what has been already done. For this reason, it seems as if we were constantly reinventing MT and both the techniques and the resulting translations are often very similar to those of ten or even twenty years ago.

Fortunately, in the last decade, propelled by the globalization of the Internet, open-source strategies have established as a sound development practice allowing for reuse of code and data. Under this new situation, developers can now focus on improving and extending available software and data. In order to ease collaborative development, open-source projects are managed on centralized websites which also act as source code repositories. Another fundamental aspect for open-source projects to succeed is the availability of complete documentation describing it.

In the last years, open-source programs and data have also appeared in the field of MT, coexisting with commercial alternatives and bringing new opportunities which are proving very positive on both research and business areas. In this paper we present a real case of these positive effects achieved by Opentrad.

In the Opentrad project two different but coordinated designs have been carried out. The differences are due to the distance between the languages:

- An open-source shallow-transfer MT engine for the Romance languages of Spain (the main ones being Spanish, Catalan and Galician).
- A deeper-transfer engine for the Spanish Đ Basque pair.

Some of the components (modules, data formats and compilers) from the first architecture will also be useful for the second. Indeed, an important additional goal of this work is testing which modules from the first architecture can be integrated into deeper-transfer architectures for more difficult language pairs.

An overview of two translation method architectures are presented in section 2; section 3 explains available languages for Opentrad; section 4 explains the innovative part in Opentrad; section 5 summarizes real cases and possible

scenarios to apply MT. Finally, section 6 ends the paper with a brief discussion.

2. System architecture

In this section we will describe the two architectures used, Apertium for related-language pairs and Matxin for the Spanish-Basque pair.

2.1. Apertium

In this section we briefly describe Apertium (Armentano-Oller et al. 2006; Corbí-Bellot et al. 2005), an open-source shallow-transfer MT engine, initially intended for related-language pairs (such as Spanish±Catalan, Spanish±Galician, Spanish± Portuguese, Czech±Slovak, Swedish±Danish, Kirwanda± Kiswahili, Bahasa Indonesia±Bahasa Melayu, etc.), but being currently extended to translate between not so related languages (such as Spanish±French and Catalan±English); an early version of this extension is expected to be released by the end of 2006. Apertium's engine, linguistic data, and documentation can be found at the project's website at http://apertium.sourceforge.net.

The open-source MT architecture Apertium is mostly based upon that of systems already developed by the Transducens group at the Universitat d'Alacant, such as the Spanish±Catalan MT system interNOSTRUM (Canals-Marote et al. 2001), and the Spanish±Portuguese translator Traductor Universia (Garrido-Alenda et al. 2004). Both systems are not open-source; however, interNOSTRUM is publicly accessible through the net and used on a daily basis by thousands of users; Traductor Universia was also publicly accessible for some years until it was converted into a full commercial product.

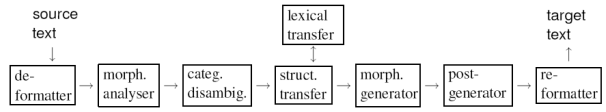

Figure 1: *The eight open-source modules of the Apertium MT system.*

The Apertium MT engine is a classical shallow-transfer or transformer system consisting of the following pipelined modules (see figure 1):

- A de-formatter which separates the text to be translated from the format information (RTF and HTML tags, white spaces, etc.). Format information is encapsulated so that the rest of the modules treat it as blanks between words.
- A morphological analyzer which tokenizes the text in surface forms and delivers, for each surface form, one or more lexical forms consisting of lemma, lexical category and morphological inflection information.
- A part-of-speech tagger which chooses, using a first-order hidden Markov model (Cutting et al. 1992) (HMM), one of the lexical forms corresponding to an ambiguous surface form; this is the only statistical-centered module.
- A lexical transfer module which reads each source-language lexical form and delivers the corresponding target-language lexical form by looking it up in a bilingual dictionary.
- A structural transfer module (parallel to the lexical transfer) which uses a finite-state chunker to detect patterns of lexical forms which need to be processed for word reorderings, agreement, etc. and performs these operations.
- A morphological generator which delivers a target-language surface form for each target language lexical form, by suitably inflecting it.
- A post-generator which performs orthographic operations such as contractions (e.g. Spanish del=de+el) and apostrophations (e.g. Catalan l'institut=el+institut).
- A re-formatter which restores the format information encapsulated by the de-formatter into the translated text.

The modules of the system communicate to each other by using text streams, which allows for easy diagnosis and independent testing. Furthermore, some modules can be used in isolation, independently from the rest of the MT engine, for other natural-language processing tasks. This extra application is also possible thanks to the full separation (or decoupling) of code and data.

2.2. Matxin

The engine is a classical transfer system consisting of 3 main components: analysis of Spanish, transfer from Spanish to Basque and generation of the Basque output.

It is based on the previous work of IXA Taldea (Díaz de Ilarraza et al., 2000) but with new features and a new aim: interoperability with other linguistic resources and convergence with the other engines in the Opentrad project through the use of XML. The previous object-oriented architecture is turning into an open-source one. This way we will be able to use modules which are shared with other engines in the Opentrad project and will comply with its format specifications.

The main modules are five: de-formatter, Spanish analysis based on FreeLing (Carreras et al., 2004), Spanish-Basque transfer, Basque generation and re-formatter.

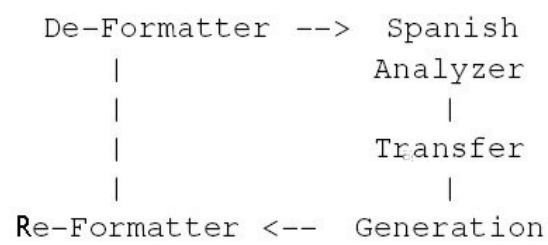

Figure 2: *Matxin MT system's architecture.*

The transfer and generation phases work in three levels: lexical form (tagged as node), chunk and sentence.

No semantic disambiguation is applied, but a large number of multi-word units representing collocations, named-entities and complex terms are being included in the bilingual dictionary in order to minimize this limitation.

2.2.1. Hybridization

In order to improve obtained results with deep-transfer methods we have some projects to develop a hybrid MT system. Currently we have restricted the linguistic field to administrative language and we are applying the following method:

- Firstly the system strips phrases from the text.
- We look for phrases in an example-based MT system and if it finds a match, it will translate the phrase.
- When there is no match, we translate phrases using a statistical MT system. To do so, we only validate sentences that reach a given threshold.
- Finally, if it does not reach that threshold, we will translate the phrase with a deep-transfer method.

We expect to get the results of this hybridization system in early 2008.

2.3. Linguistic data

Apertium's linguistic data (which are fully decoupled from the translation engine) are coded using XML-based formats; this allows for interoperability (that is, the possibility of using the XML data in a set of different scenarios) and for easy data transformation and maintenance. On the one hand, the success of the open-source MT engine heavily depends on the acceptance of these formats by other groups; this is indeed the mechanism by which *de facto* standards appear. Acceptance may be eased by the use of an interoperable XML-based format, and also by the availability of tools to manage linguistic data. But, on the other hand, acceptance of the formats also depends on the success of the translation engine itself. The XML formats for each type of linguistic data are defined through conveniently-designed XML document-type definitions (DTDs) which may be found in the Apertium and Matxin packages.

There are four sets of linguistic data organized at two levels: lexical or morphological level and structural or syntactical level:

- At lexical level morphological and bilingual dictionaries are used following the proposal for the whole Opentrad project.
- At structural level two grammars are being developed: one for structural transfer and other for syntactical generation.

3. Available language pairs

As we see in the section above, we have two translation engines (Matxin and Apertium). However, right now Matxin only works with the Basque-Spanish language pair so we will focus on Apertium.

Apertium's MT engine has been released in two open-source packages: *lttoolbox* (containing all the lexical processing modules and tools) and *apertium* itself (containing the rest of the engine); both are available under GNU GPL license. In addition to these programs, open-source linguistic data are already available for various language pairs:

- The Spanish±Catalan (packaged under the name apertium-es-ca) and Spanish±Galician (package apertium-es-gl) pairs developed under the Opentrad consortium and released under GNU GPL license;
- The Spanish-Portuguese pair (package apertium-es-pt) developed at the Universitat d'Alacant and released under GNU GPL includes the Brazilian variant as well;
- Pilot data for Catalan±French (package apertium-fr-ca) and Catalan±Occitan/Aranese (package apertium-oc-ca) released under GNU GPL;
- Pilot Catalan±English data.
- Spanish±French pair, expected to be released by the end of 2007.
- Pilot for Spanish-English, expected to be released by the end of 2008.

Apertium gives a reasonably good translation quality between related languages (error rates around 5-10 percent in general purpose translations). These results are obtained with the pilot open-source linguistic data already released (having around 10,000 lemmas and less than 80 shallow transfer rules) which might easily improve mainly through lexical contributions from the linguistic communities involved. The Apertium open-source engine itself is being actively developed and contributions to its design may enhance it to perform more advanced lexical and structural processing tasks.

All the available packages and documentation for Apertium are hosted at http://www.sourceforge.net/projects/apertium. Additional information may be found at http://www.apertium.org. Finally, web prototypes for MT systems for all the currently available pairs may be tested on plain texts, RTF and HTML at http://xixona.dlsi.ua.es/prototype.

4. Innovation

Some of the most important innovations of the Opentrad project is its open-source development methodology and introduction of a new business model in MT field.

4.1. Methodology

Open-source software projects are based on collaboration. This way, the source code of these projects is available on the net so that anyone can participate on the development. Many projects are hosted on personal websites but there are other hosting alternatives presenting many advantages such as control of versions, increase of visibility or developers management. These websites allow for centralized collaboration and distribution, and are known as open-source development websites. Two of the most commonly used are:

- SourceForge.net is the world's largest open-source software development website, hosting more than 100, 000 projects and over 1,000,000 registered users with a centralized resource for managing projects, issues, communications, and code.
- Savannah is a central point for development, distribution and maintenance of open-source software that runs on free operating systems. It hosts more than 2,500 projects and has over 45,000 registered users. It includes issue tracking, project member management by roles and individual account maintenance.

The two Opentrad MT engines are hosted at SourceForge.net and their web sites are: *apertium.sourceforge.net* for Apertium and *matxin.sourceforge.net* for Matxin.

As Opentrad is an open-source project, all software developed can be downloaded. Apart from that, anyone can see its code, change and publish it. This gives new opportunities for research in different MT fields and from the other side, it also gives new opportunities for business as we will explain in the next section.

4.2. Business model

It is not easy to encourage clients to use open-source software: open and free terms are usually perceived as untrustworthy. However, there is a strong reason why clients would prefer open-source to closed-source software: clients who choose open-source software do not see companies distributing them as providers to whom they have a technological dependency, but as technological partners, since clients may feel free to contract services around the open-source system with any other company offering them; therefore, technological dependence, a typical feature associated with closed-source products, is strongly diminished.

Even more interesting for institutions, public entities and large companies is the social action they can contribute to by making open modifications, improving data or adding new functionalities to the open-source software specially developed for them. This gives them a very positive image before their clients and users; they are not only offering a better service, but also benefiting the whole community.

It is also very difficult to convince tech companies to make their software open-source. The point is the change of the business model from a product-selling centered model to a service-offering one. Innovative services around a good open-source software are the main competitive advantages of this business model. Besides, contributions to the open project coming from elsewhere are also contributions that companies can benefit from in order to offer better products and services. This non-controllable aspect of the development makes heavy demands on those companies offering services based on open-source software but, despite this effort, in the current world it is crucial for tech companies to remain constantly updated.

Open-source software pose business challenges for those researchers working on new methods and techniques. Indeed, the number of technological-based spin-offs (here, companies created by researches as a result of a particular research activity) has increased in the last few years.

These companies have not only a product and a catalog of related services to offer, but also the know-how developed during the research work of their members, and, being half-way, they can offer the best services in collaboration with universities and companies.

5. Business real cases and possible scenarios

Open-source software also brings new business models to private companies. Taking the Apertium MT system as an example, companies can offer a wide variety of services around it, such as these:

- installing and supporting translation servers;
- maintaining, adapting and extending linguistic data;
- building data for new language pairs;
- integrating MT systems in multilingual documentation management systems
- Offering full translation services based on MT
- developing new tools for Apertium, etc.

Furthermore, companies and individual translators can adapt linguistic data to restricted language domains or to dialectical varieties in order to ease post-edition or better suit their clients' needs when offering translation services.

An illustrative example of companies benefiting from some of the previous profitable market segments is the case of the three companies (Eleka Ingeniaritza Linguistikoa, Imaxin| Software and Elhuyar Fundazioa) participating in the

Apertium project as well as the case of a new company named Prompsit Language Engineering, created to exploit the challenges derived from the existence of Apertium.

5.1. Newspaper on-line edition translation

Imaxin software has improved and adapted the initial translation-engine with the Spanish-Galician pair, for the journal ªLa Voz de Galiciaº, the eighth most-read newspaper in Spain. The process to have this machine translation system integrated into the journal editing environment took six months. From then on, the journal's on-line version is available both in Spanish and in Galician. It achieves less than a 5% error rate which results in a good translation quality.

Imaxin has also developed a tool to manage dictionaries for the machine translator so that new words can be easily added into the translator. At this moment, human review is needed in order to eliminate this error rate.

Finally, La Voz de Galicia has decided to leave the dictionaries resulting from this project free. That means that anyone can use the currently largest Galician dictionary.

5.2. Automated full translation service

Elhuyar Fundazioa offers full translation services through Apertium. Translation services are one of the commercial services offered by this foundation. Elhuyar uses Apertium to offer full translation services based on MT: translation using Apertium, linguistic correction and terminology services.

The use of Machine Translation systems integrated on commercial linguistic services gives commercial advantages to a translation service company as Elhuyar:

- Possibility to offer competitive prices for MT language-pairs maintaining linguistic quality
- Specialization in language review instead of human translations
- Integrated translation services: thanks to the best rates in MT language-pairs, we can offer human translation services at a lower price and therefore attract new customers.
- Terminology management services

5.3. Enterprises in internationalization process

In this world of globalization many companies are internationalizing their business, not only spreading their sales areas but also setting new production-plants in countries like the Czech Republic, Romania, Poland or Brazil. In this process enterprises may need to solve communication problems with employees and may also have to integrate in those countries' culture and language.

MT can help totally in this process; human-translation is slow and high-cost to translate piles of documentation generated in the enterprise, while MT systems can translate them fastly. Being open-source anyone who has linguistic and MT knowledge can adapt the system to each organization.

This is one interesting business area because languages like Portuguese (Brazilian) can be translated with very good accuracy.

5.4. Tourism industry

Other interesting scenario is the industry of tourism. Public organizations of tourism do a great effort to offer touristic information in several languages. In the Basque Country for example, they have a web site in Spanish, Basque, English and French. But a lot of tourists come from Catalonia, so to give a better service they can use MT to translate their web page into Catalan. An integration of an open-source MT system to translate web pages is not too difficult and the main task to do would be an adaptation of the dictionaries to the local toponymy, very rich in tourism-related texts.

Local and regional governments are interested in solutions like this to reduce translation costs, but the main obstacle is to find funding opportunities to cover the adaptation of dictionaries.

5.5. Other

Also, some institutions and public entities are juggling with the possibility of installing Apertium as their MT platform to offer on-line translation services. Banks which are present in different heterogeneous linguistic areas have also shown their interest in integrating Apertium-based MT systems in their documentation management systems.

6. Discussion

With the recent trends in open-source software development, new challenges raise for both research institutions and companies. Open-source practices have recently reached the MT arena, therefore introducing new perspectives on MT system development. A new business model, which focuses on the services around translation engines and linguistic data more than on the programs and data themselves, is possible.

In this paper we have presented Apertium, a full open-source MT system with a lot of potentials and introduced the main aspects around the new business model inspired by the Apertium system and the current state of development of an open transfer MT architecture for Spanish-Basque.

We have also presented different areas of business where open-source MT systems and translation-service providers can be very interesting allies.

7. Acknowledgements

Work funded by the Spanish Ministry of Industry, Commerce and Tourism through project Opentrad (FIT-340101-2004-3,

FIT-340001-2005-2, FIT-350401-2006-5, FIT-350401-2007-1) and Basque Government Department of Industry, Commerce and Tourism through project Opentrad (GAITEK IG-2006/00371)

8. References

[1] Corbí-Bellot, A. M., Forcada, M. L., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Alegria, I., Mayor, A., and Sarasola, K. (2005). An open-source shallow-transfer machine translation engine for the Romance languages of Spain. In Proceedings of the 10th European Association for Machine Translation Conference, pages 79±86, Budapest, Hungary.

[2] Alegria I., A. Diaz de Ilarraza, G. Labaka, M. Lersundi. A. Mayor, K. Sarasola, (2005) A FST grammar for verb chain transfer in a Spanish-Basque MT System. Proc. of the Finite State Methods in Natural Language Processing workshop. Helsinki.

[3] Carreras, X., I. Chao, L. Padró and M. Padró (2004). FreeLing: An open source Suite of Language Analyzers, in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal.

[4] Garrido-Alenda, A., Gilabert Zarco, P., Pérez-Ortiz, J. A., Pertusa-Ibáñez, A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M. A., and Forcada, M. L. (2004). Shallow parsing for Portuguese-Spanish machine translation. In Branco, A., Mendes, A., and Ribeiro, R., editors, Language technology for Portuguese: shallow processing tools and resources, pages 135±144. Lisbon.

[5] Díaz de Ilarraza, A., A. Mayor, K. Sarasola (2000). Reusability of wide-coverage linguistic resources in the construction of a multilingual machine translation system, in Proceedings of MT 2000 (Univ. of Exeter, UK, 1922 Nov. 2000), .

[6] Forcada, M. L. (2006). Open-source machine translation: an opportunity for minor languages. In Proceedings of Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages).

[7] World Wide Web Consortium (2004). "Extensible Markup Language (XML)", http://www.w3.org/XML/.