Building Resources for Machine Translation:

What the user hasn't got we have to provide.

Gudrún Magnúsdóttir

Språkdata
University of Gothenburg
S-412 98 Gothenburg

Team Zilos
Järntorget 2
S-413 04 Gothenburg

email: gudrun@svenska.gu.se

1. *Introduction*

The greatest sources of language data for natural language processing are held by the machine translation development community. That data is potentially more in demand than the MT-systems themselves. The defensive attitude of not making these data available for further development is damaging the natural evolution in the field.

Activation generates users and those in turn the number of systems to be bought. However, that activation is stalled primarily by the cost of building an MT-system, i.e. the lack of language data available, and secondly by the fact that the potential buyers of machine translation systems lack the knowledge needed for tuning the system to fit the in-house environment.

The first factor limiting activation can be attacked from within the MT-community namely by making the language data available to potential competition at a price that is beneficial to both parties. The enhancement of the data will be well taken care of by further MT-development as well as that of building general language engineering resources within the European Community. The issue of re-usability of the data is directly related to linguistic purism that we will not discuss further.

The second factor is that of human resources and has usually been regarded as something an expensive evaluation will solve. However, that doesn't solve the problem. The potential buyer needs a system solution based on analyzed need and after the purchase of any MT-system it needs to be tuned to the site environment. The buyer/user usually hasn't got the knowledge to go beyond that of identifying a need, the selection of a system as well as the in-house tuning of the system is thus ad hoc.

2. *Usable Language Resources*

Current initiatives in the Language Engineering programme ensure continued enhancement of resources. However, these enhancements do not include bilingual-lexicons needed for machine translation and perhaps they shouldn't. There may be a case for theoretical improvement at transfer/interlingua level in MT, whereas the analysis as such is stagnated in a heavily lexicalised solution. This is followed by enhancement of lexical semantics, in MT as well as other applications, which may hopefully open new possibilities for linking the translation. It should not be argued that the rule-based solutions with heavy semantics is the only way to move forward. In the past few years translation memories applying advanced pattern matching have proved their relevance and we have not as yet seen the real implications of how these two methods can be successfully linked.

However important the Language Engineering (LE) initiatives for language resources are, they are hampered by the need for standardization and would as such on the bilingual/multilingual level possibly hinder theoretical development at the transfer/interlingua level.

Several MT-systems have impressive lexicons and grammars, the coverage of which cannot matched by any other application in LE. Most of the lexicons have been developed for decades and as such been tuned to special requirements. This doesn't imply that the systems themselves have been improved even if their performance has. The kernel in most MT-systems is regrettably ancient history whereas their lexicons and linguistics are positively gold for any LE-developer. The tricky bit is getting it out of there without the roof falling down on your head.

The typical MT-system lexicon will in fact give you the data needed for basic natural-language analysis. The transfer component will to some degree give you data to do any multilingual application, such as information retrieval across language-pairs. The usage of the data is linked to the willingness to use it as well as the availability. Hopefully MT-developers have become mature enough to realize that their systems as they are will not bring them fame and fortune and the reverse, those, that have fervently argued that everything must be developed from scratch, since linguistics in MT-systems aren't according to their favourite theory, will accept that it is better to perform than not at all.

I would also like to mention that the usage of MT-systems beyond MT has not been explored in detail. It has been argued that the best text to align is MT translated text. Perhaps machine translation focussing on document management is a very dated application that carries within the resources for several new offsprings.

3. Human Resources

It is clear that the marketing of MT is a failure. The market need and the marketed systems never meet and once they do a lovers quarrel is guarantied. Without the market, MT is dead as an application. However, as long as there is a need there is a market and it is simply a matter of getting there without pushing it down anyone's throat nor running away at the sight of a customer.

The communication between the MT-sales people and the buyer can in the best of worlds, where the sales person knows that the system he sells translates and he is able to run it as a demonstration, and the buyer has understood that this thing does what his translator does but *not all the way,* result in a sale of a system that will be put on the shelf in three months by the poor translator that it was going to do most of the job for.

Two factors can amend this, education or/and consulting. Education should start with those currently working on sales. The consulting is a difficult issue since most of the expertise is within the development teams and thus the consultant is possibly bias and cannot be taken seriously by a buyer. It would be ideal if each company doing any type of translation had an MT-expert of its own. Currently however, the link between the identification of a need and the selection of a system solution is probably better served by any type of independent consulting.

The goal is to ensure that the buyer doesn't make mistakes that might in turn be harmful for the reputation of the field as a whole. In the past this has led to an unwanted stagnation in system sales and the question remains if more systems will be developed unless the market starts to expand.

One thing leads to the another, more users of MT-systems enhance the language data within the systems as well as provide funding for improved systems development.

4. Conclusion

In the light of this, human resources are potentially the most important factor, ensuring continued activation in machine translation by providing the buyer with the knowledge that he lacks.

References.

(1987) Magnusdottir, G. FASTCAT, Machine Translation Systems and Translators Interviews, EUREKA Project report.

(1993) Magnusdottir et al. The Swedish Eurolang Report, NUTEK (The Swedish Board of Industrial and Technical Development) Project Report.

(1994) Magnusdottir, G. Multilingual Trademark Retrieval and Access. (Classified)