Localizing Content Management Application for Spontaneous Textual Utterances in Natural Language.

Najeh Hajlaoui

Localizing content management application for spontaneous textual utterances in natural language

NAJEH HAJLAOUI

LIG Laboratory, University Joseph Fourier,
385 rue de la Bibliothèque, BP n° 53
38041 Grenoble, cedex 9, France
33+ (0)4 76 51 43 79

Najeh.hajlaoui@imag.fr

Abstract

The multilinguisation of content management services is an important but difficult problem and very few services do it. In fact it depends on the translation situation: types and level of possible accesses, available resources, linguistic competences of participants in the multilinguisation of application. Several strategies of multilinguisation are then possible (by translation, by internal or external localization etc.). We illustrate this study by a real case of linguistic porting (Arab to French) of an E-commerce application deployed in Jordan, using texts of spontaneous SMS for buying and selling second-hand cars. In spite the long distance between Arabic and French, the localization methods used give good results because of the proximity of the two sublanguages of Arabic and French.

Keywords*: Localization, spontaneous texts, Content Extraction, SMS (Shorts Messages Services), CRL (Content Language Representation).*

Introduction

Multilinguisation of E-commerce services using spontaneous texts in natural language is an important but difficult problem, and very few services do it. It depends on two principal factors on the translation situation:

- The access level to resources of the initial application, four cases are possible: complete access to the source code, access limited to the internal representation, access limited to the dictionary, and any access.
- The linguistic qualification level of participants in the multilinguisation of the application, which can be defined by knowing or not the source language, or by the linguistic abilities of the team in charge for the localization (specialist or not in NLP).

We present primarily the requirement in localization of content extraction application followed by an analysis for the possible methods. Then, we illustrate our study by the linguistic porting (from Arabic into French) of an E-commerce application for which the two factors presented above are assured, and several strategies of multilinguisation are then possible. We present a strategy of localization called "external" which requires a simple access to the internal representation. Then, we evaluate it results compared to the results given by the original version system and by another localization method called "internal", which consists adapting the existing content extractor. This internal method is described in an other paper submitted.

1 Multilinguisation of content management services is a important and difficult problem and very few services do it

1.1 Requirement in localization for content extractor in natural language

The characteristics of the applications, which we are interested in primarily, are those of the E-commerce, and those that give added value by the processing of the content. Main types of applications and services are : categorization of various documents (messages of AFP "Agence France-presse", messages of customers on a SAS "Service After Sale" server,…) : "Who is interested by what", information extraction to feed or consult a data base (example: small ads, FAQ "**F**requently **A**sked **Q**uestions", automated hotlines).

1.2 Increasing importance of offering multilingual services

Offering a multilingual service is a necessity for deployment in many countries. For example, CISCO society (http:// www.cisco.com/) is obliged to use translation systems to (badly) translate its documentation. Indeed, the call centres are overloaded, and they are very expensive. This example shows the necessity of local multilingual services : of course, in multilingual countries (Canada, India, US), but also in monolingual countries (France) because of tourism and new mobility.

1.3 In spite of the interest of spontaneous utterances processing, few services do it

1.3.1 Interest

The interest of a spontaneous utterances processing service is summarized in the following points:

- Naturalness of interaction,
- Possibility for the users of expressing anything, whereas forums and menus, in modal character, do not allow it,
- Possibility (more recently) of building and let evolve the design domain-focused service (thesaurus ontology job).

A service of spontaneous texts processing (ads, messages…) can be an answer to a technical and ergonomic need, especially in the E-commerce domain.

Menu-driven navigation and keyword search provided by most commercial sites have tremendous limitations, as they tend to overwhelm and frustrate users with lengthy and rigid interactions. User interest in a particular site decreases exponentially with the increase in the number of mouse clicks (Huberman, Pirolli et al. 1998). Hence shortening the interaction path to provide useful information becomes important. Many e-commerce sites attempt to solve the problem by providing keyword search capabilities. However, keyword search engines usually require that users know domain specific jargon, so that the keywords could possibly match indexing terms used in the product catalog or documents. Keyword search does not allow users to precisely *describe their intentions, and, more important, it lacks understanding of the semantic meaning of the search words and phrases.*

1.3.2 Few services treat spontaneous texts, even in monolingual context

After having interrogated several search engines (Google, AltaVista, Tiscali…) on the subject of processing spontaneous texts in natural language applications, we obtained very few positive results. When we found something, we had very few information about internal procedure and their multilinguisation.
We passed as request, several and various keywords (localizing natural language message processors, localization NLP free text, localization NLP interfaces, multilingual customer message processsing, multilingual customer messages tools, multilingual customer relationship processing, multilingual NLP e-commerce, multilingual online sales customer support, multilingual online sales NLP customer support, categorizing natural language message, handling natural language messages in business, Natural Language Conversational Interface in Online Sales…).

It seems that there are still very few ! Few applications or theories correspond even partially to our request, like Pertinence Summarizer (Lehman 1996) - a software of automatic summary of multilingual texts -, Amilcare (Ciravegna 2001) - an adaptive system of information extraction -, NLSA "Natural Language Sales Assistant" - a dialogue-based system through the Web deployed by IBM - and still better CATS "Classifieds Ads Through SMS" (Daoud 2006) - a Arabic-based SMS system for buying and selling cars and real estate.

1.3.3 Possible reasons (why few services treat the spontaneous texts in natural language)

Our hypothesis on possible reasons for which very few services process spontaneous texts in natural language is that this work has inherent difficulties and that their multilinguisation creates large problems. These are similar to problems encountered in speech processing: "not standard" grammar (more or less oral), errors (spelling, typing errors), use of typographical conventions suitable for the context (ex SMS, chat, emotion, abbreviation). Often enough, we are in a "sublanguage" relatively far from the "general language", as small ads or alarms/warnings (road traffic, natural disasters).
In general, we can not use tools made for written "general" and "clean" language. It is necessary to "translate" spontaneous texts in natural language into a "content representation " formalism and in most cases, an application has its own formalism.

1.4 Necessity of a specific approach to each sublanguage and inefficiency of tools made for the general languages

Applications processing spontaneous texts in natural language generally use a content representation. We find several forms of content representations like lists <attributes, values>, structures of typed features, logical expressions (Prolog), logico-functional expressions, objects (classes (methods, attributes), instances). For example, the CATS system uses a representation like Propriety = colour {object = saloon, value = blue } for selling and buying second-hand cars. Often, an application has its own formalism. Adapting or porting a "content extractor" from one application to another is difficult, even for the same language, because it is necessary to be able to guarantee a minimum level of quality, correctness/completion of the content extraction, relevance of produced answers (processing + management) and linguistic adequacy of these answers.

The primary data to provide for a linguistic porting of an application treating spontaneous texts in natural language is a corpus of the same type of data, which is not always obvious to find. It is generally necessary to work by simulation/imagination and adaptation. For example, CATS treats real estate SMS in Amman. But to adapt CATS into French supposes to choose a city in France, Belgium, Switzerland, Canada, or Africa… And when we want to port into languages unknown to the developers and owners of the application/service, that becomes very expensive.

Generally, the "sublanguage" used (naturally) is relatively far from the "general language". Consequently, the existing tools (Machine Translation or Content Extraction tools) made for the "general language" do not work. Under these conditions, specific approach to each sublanguage is then essential.

2 Reflexions about possible approaches

2.1 By translation

A first idea consists in translating spontaneous texts from L2 language into the original language L1 of the application to be located. The translation and the revision of an existing corpus then the adaptation to the corresponding sublanguage in the other language, can help the localization operation and especially enable to create an initial version of corpora, but the automatic translation systems are not available for all language pairs. Very far from there, they are useful only for "clean" texts and does not work with spontaneous texts. Adding the construction of a targets corpus problem to allow the evaluation of the porting operation. As an example, for artisanal development, we can take training words, but difficulty which arises moreover, it is how to define the "good size of corpus". The idea of realisation of an automatic translation system for spontaneous texts is not impossible but several questions arise: which type of system? Which size of resources must we have?

2.1.1 Creation a statistical system

For the case of the statistical systems, two main problems have to be solved:

- Estimation of the bilingual corpus size necessary according to the estimated "proximity" of the native and target sublanguages.
- Construction of this corpus.

The experiment can start with a minimal size under the effect of the sublanguages proximity.

2.1.2 Creation a symbolic system (dictionaries, rules, heuristic)

The advantage of the creating a symbolic system based on heuristic, dictionaries, and rules is that it can be very small and developed in a few weeks. On the other hand, the major disadvantage is that it is necessary to have a linguist expert specialist of the concerned language.

2.2 By realization of a new content extractor for each concerned language

The realization of new content extractor can be done by several methods.

The first solution is to adapt the existing content extractor from L1 to L2, but that is viable, only if

- the developers agree to open their code/tools to cooperate with collaborators
- This code/tools is also easy to understand
- The resources are not too heavy to create (in particular the dictionary if there exists)
- Maintenance then can be done at simple cost, by simple collaborations.

This method of "internal localization" of native content extractor requires of course a training of the localization team with tools and methods used.

The second solution would consist, for a company wanting to offer multilinguisation/porting services. The idea is to implement a generic content extractor and to adapt it to each situation (languages, sublanguage, domains, CRL "Content Represntation Language, task, constraints). We will see in the following part that seems very difficult to consider currently.

In several contexts, "contributors" in multilinguisation task haven't access to application content extractor, nor to an "universal" content extractor. In this case, a third solution could then be to find and to adapt an existing and available content extractor, it can be for the same concerned language L2, with a different domain/task, or it can be for another language (different to L2) but with the same domain/task.

We illustrate the first method by a real case of linguistic porting (Arabic to French) of CATS system by adapting their content extractor, initially envisaged for Arabic SMS. This method is called "internal adaptation" and it is described in other paper to be published.

We illustrate the third method by a real case of linguistic porting (Arabic to French) of CATS system by an "external adaptation". It consist to adapt another existing content extractor and to translate the results in the CATS content representation formalism. This method requires a simple access to the internal representation. We describe this method in the following section.

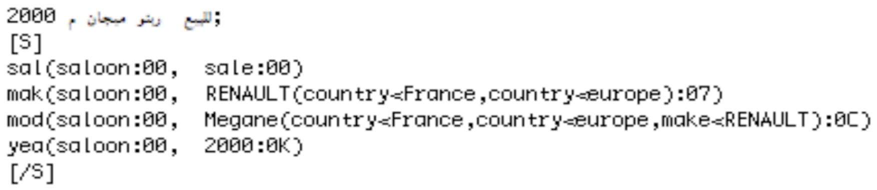
3 Illustration: localization of the CATS system (Arabic-based to French)

3.1 Presentation of CATS

CATS (Classifieds Ads Through SMS) is a platform for buying and selling goods (cars, real estate…) based on the use of the Arabic SMS and created by Daoud. It is deployed by Fastlink, the largest mobile phone operator in Jordan. Although there is no direct transaction, it helps the users to sell and to buy by putting them in contact. These SMS are sent to a single special number and are recorded automatically in a data corpus. Their texts are analysed and an answer is sent automatically to the sender of the SMS if the request corresponds to one or more proposals. If nothing is found, the system says it in its answer, and resends the request later when the data base changes.

The structure of CATS is summarized in two main components: a Content Extractor "CE" and a Query Manager "QM".

CE receives the SMS and decodes their texts into a CRL-CATS representation (Content Representation Language) by using lexicons specific to the domain. CRL-CATS is a Property {object, value} representation.



```
;للبيع  رينو ميجان م 2000
[S]
sal(saloon:00,  sale:00)
mak(saloon:00,  RENAULT(country<France,country<europe):07)
mod(saloon:00,  Megane(country<France,country<europe,make<RENAULT):0C)
yea(saloon:00,  2000:0K)
[/S]
```

;Selling Renault Megane m 2000

Figure 1: example of CRL-CATS representation with its English translation

In the preceding example, the second property is make (mak), the object is a car (saloon) and the value is equal to (RENAULT(country<France,country<europe)). For the property model (mod), the object is a car (saloon) and the value is equal to (Megane(country<France,country<europe,make< RENAULT)). For the property year (yea), the object is a car (saloon) and the value is (2000).
The QM converts the CRL-CATS representation into a SQL text (selection request for the buying and/or insertion request for selling). It treats also the situations in which no answer is found.

3.2 Necessity of starting corpus

For all localization methods of CATS system into French, the first thing to do is to collect or build a "French starting corpus", similar to that used by Daoud at the beginning of his project for Arabic. That was obviously necessary to study the syntactic form of the SMS to be treated in French, and to see also which lexical categories to expect. A first idea is to use the CATS corpus in Arabic and to translate it into "French spontaneous SMS", supposed to be sent (in Jordan) by French-speaking people.

A rough translation produced by a non-French person is generally very different compared to a natural and functional translation produced by a French person, i.e. compared to what a French person would say in a spontaneous way in the same situation. We evaluated this translation difference between rough and natural or literal and functional by calculating the edit distance between two translations. The average distance is 21,88 (Hajlaoui 2006). The distance used is the minimal number of suppressions, insertions or replacements of letters necessary to transform a rough (or a literal) translation into a natural (or a functional) translation.
We showed in a precedent paper (Hajlaoui 2006) that direct translations of a real corpus of natural sentences can not give natural results. Taking this hypothesis, we tried to produce a small French corpus functionally equivalent to the Arabic initial corpus.

In order to develop this corpus, we adopted the following technique: starting from the ads model constituted by 50 types of SMS revised and considered to be functional, we multiplied the number of these ads by forming different arguments combinations used (make, model, year, color, price…). For example, we replace a year by another (je cherche une voiture modèle 98) → (je cherche une voiture modèle 99…) or a make by another (A vendre BMW rouge) → (A vendre PEUGEOT noire)…

3.3 External localization

An external localization method consists in adapting an existing and available content extractor, for the same language (French), but to change the domain (from tourism to automobile). Concretely we adapted an existing French extractor of 31 918 lines of code in Tcltk, created by H.Blanchon for the Nespole! project (Blanchon 2004). In his work, Blanchon used the IF "Interchange Format" representation.

The IF is a semantico-pragmatic pivot used for restricted domains. In the NESPOLE! project, the passage to the IF pivot uses a method based on the relevant sequences recognition automat. The following figure shows the IF specification components: speech act, concept and arguments. In the beginning of this adaptation, we have the code of the second demonstrator, the paper and electronics version of the IF specification (version of 08-18-2002) and the CRL-CATS specification.

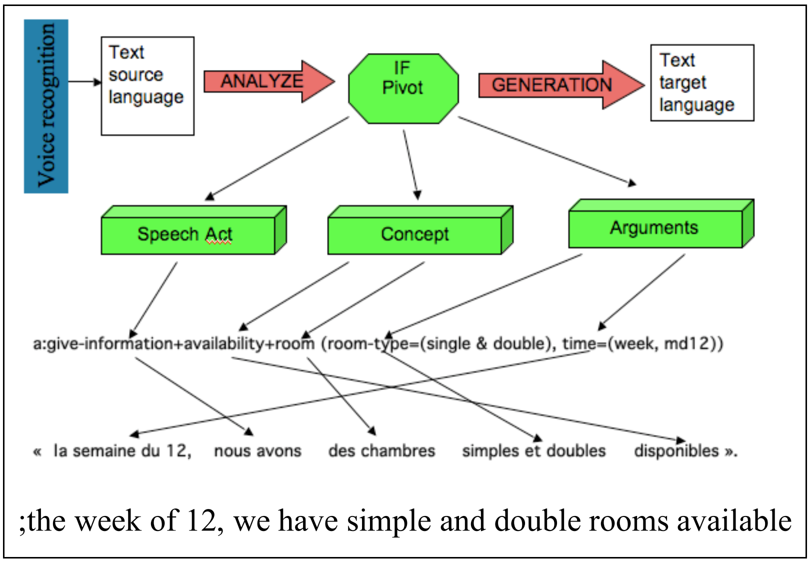

Voice recognition
Text source language
ANALYZE
IF Pivot
GENERATION
Text target language
Speech Act
Concept
Arguments
a:give-information+availability+room (room-type=(single & double), time=(week, md12))
« la semaine du 12, nous avons des chambres simples et doubles disponibles ».
;the week of 12, we have simple and double rooms available

Figure 1 : content extractor for French in tourism domain (Blanchon 2004)

3.3.1 Content extraction method in Nespole!

We tried to understand and use the method used in the second module. As the following figure shows, the method used for the analysis (French to IF) has the following stages:

- Segmentation in SDU (Semantic Dialogue Units).
- Detection of the domain.
- Construction of speech acts prefix and instantiation of dependent arguments.
- Instantiation of arguments related to domain and management of subordinations.
- Complementation of speech acts

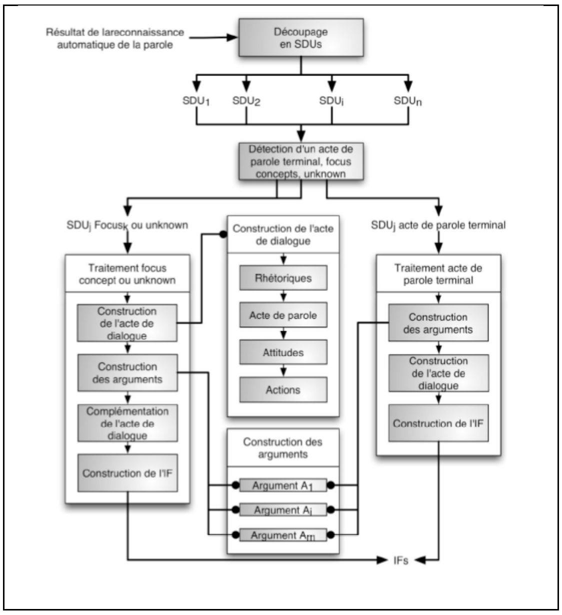

Résultat de lareconnaissance automatique de la parole
Découpage en SDUs
SDU_1
SDU_2
SDU_i
SDU_n
Détection d'un acte de parole terminal, focus concepts, unknown
SDU_j $Focus_k$ ou unknown
Construction de l'acte de dialogue
SDU_j acte de parole terminal
Traitement focus concept ou unknown
Construction de l'acte de dialogue
Construction des arguments
Complémentation de l'acte de dialogue
Construction de l'IF
Rhétoriques
Acte de parole
Attitudes
Actions
Construction des arguments
Argument A_1
Argument A_i
Argument A_m
Traitement acte de parole terminal
Construction des arguments
Construction de l'acte de dialogue
Construction de l'IF
IFs

Figure 2 : Structure of French analysis module into IF
for the NESPOLE second demonstrator (Blanchon 2004)

3.3.2 Adaptation of the IF into IF-CATS and of the French-IF code

We adapted IF specification to the cars domain. We also enriched it by adding new arguments like vehicle-motor-type, vehicle- hand… We added new actions, essentially buying action e-buy and selling action e-sell. We used the same stages to extract information about the cars domain. We tried to eliminate the instructions which posed problem and/or which were not necessary to reduce the computing time. We added new instructions related the added arguments and actions.

Much work was done in the arguments instantiation stage related to the “vehicle” domain: we instantiate essentially vehicle specification “vehicle-spec”, as well as other less interesting arguments such as: theDistance, theLocation, theDuration, theDestination, theTime, thePrice…

A VehicleSpec2If function allows research and construction of the arguments related to “vehicle” concept. The only argument which was programmed in the second demonstrator code is “frenchvehicle”, which can have value as “voiture, ski, camion, bus, train, avion… ”

Other arguments exist in initial IF specification, but they are not programmed, such as: “makevehicle, modelvehicle, sizevehicle, frenchcolor, agevehicle, pricevehicle”

In the same way, the Argument2if function builds the IF associated values. The following figure is an example of result obtained after adaptation.

Entrée 1 = A vendre une grande voiture française BM 325 4 portes diesel bleue TBE première main assurance complète avec CT sans climatisation TB prix dernier mod

English translation : For selling a big French car BM 325 4 doors diesel blue TBE first hand insurance completed with CT without air-condition TB price last mod

Sortie1 = {c:give-information+disposition+vehicle(disposition=(desire, who=i), action=e_sell, vehicle-spec=(car, vehicle-make=BMW, vehicle-model=325, vehicle-size=4 door, vehicle-shape=big, vehicle-motor-type=diesel, vehicle-hand=first_hand, vehicle-color=blue, vehicle-condition=good, vehicle-assurance=inssured, vehicle-controle=total_check, vehicle-air-condition=no_air_condition, vehicle-nationality=french, age-vehicle=new_mod, price-vehicle=good_price))}

Figure 3 : example of result of the content extractor for French SMS

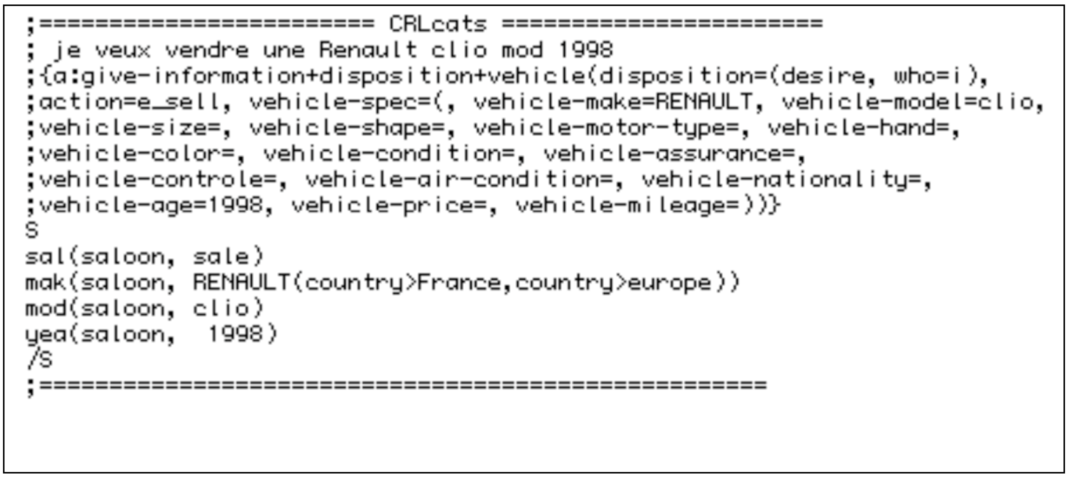
We call the obtained result "IF-CATS" (sortie1 in the preceding example).

3.3.3 Compiler IF-CATS_CRL-CATS

We built a compiler, which analyzes and transforms the IF-CATS result in the CRL-CATS format by using a "IF-CRL" dictionary related to this structure, which allows the substitution of the arguments. The following figure gives an example of the result of the compiler. It shows that it is possible to obtain the same CRL-CATS format as that produced by the EnCo[1] tool, except for the symbols 00, 0J, 0R which are added by the EnCo tool (Figure 1 is an example produced by EnCo).

```
; I want to sell a Renault clio mod 1998
```

[1] EnCo is a tool based on rules and dictionaries used for content extraction in originale version of CATS system.



```
;====================== CRLcats ======================
; je veux vendre une Renault clio mod 1998
;{a:give-information+disposition+vehicle(disposition=(desire, who=i ),
;action=e_sell, vehicle-spec=(, vehicle-make=RENAULT, vehicle-model=clio,
;vehicle-size=, vehicle-shape=, vehicle-motor-type=, vehicle-hand=,
;vehicle-color=, vehicle-condition=, vehicle-assurance=,
;vehicle-controle=, vehicle-air-condition=, vehicle-nationality=,
;vehicle-age=1998, vehicle-price=, vehicle-mileage= ))}
S
sal(saloon, sale)
mak(saloon, RENAULT(country>France,country>europe))
mod(saloon, clio)
yea(saloon,  1998)
/S
;==================================================
```

Figure 4 : an example of IF-CATS_CRL-CATS compiler result

3.4 Results and evaluation compared to the original version

3.4.1 Evaluation method

We translated manually the evaluation corpus used for the evaluation of CATS Arabic version (original). It contains 200 real SMS (100 SMS to buy + 100 SMS to sale) posted by real users in Jordan. We spent 289 mn to translate the 200 Arabic SMS (2082 words is equivalent to 10 words/SMS, approximately 8 standard pages[2]) into a French translation or about 35 mn per page. We spent 10 mn per standard page to pass from raw translation to functional translation. We obtained 200 French SMS considered to be functional (1361 words, or about 6,8 words/SMS, approximately 5 standard pages).

We computed the recall R, the precision P and the F-measure F for each most important property as follows (action “sale or buy”, “Make”, “Model”, “year”, “price”):

P = number of correct entities identified by the system/ total number of entities identified by the system;

R = number of correct entities identified by the system/ total number of entities identified by the human;

$$F = 2PR/(P+R)$$

3.4.2 Results

The percentages of porting by internal adaptation (compared to the original version) vary between 95% and 100%, with an average of 98%. The percentages of porting by external adaptation (compared to the original version) vary between 46% and 99%, with an average of 77%.

[2] standard page = 250 words

Let us note that properties treating the number like price and years that make low the value of porting percentage by external adaptation, but its advantage is it requires a simple access to internal representation of the application.

	EnCoAR (original version)			EnCoFR (internal adapatation)				RegExpFR (external adapatation)			
Properties	Precision	Recall	F-measure (ExpRegFR)	Precision	Recall	F-measure (EnCoAR)	% porting	Precision	Recall	F-measure (RegExpFR)	% porting
Buy/Sale	0,956	0,970	0,963	0,994	0,855	0,919	95	1,000	0,835	0,910	95
Year	0,817	0,960	0,883	0,879	0,819	0,848	96	0,828	0,271	0,409	46
Price	0,800	0,822	0,811	0,789	0,822	0,805	99	0,955	0,288	0,442	55
Make	0,978	0,963	0,970	0,961	0,961	0,961	99	0,994	0,928	0,960	99
Model	0,901	0,837	0,868	0,842	0,903	0,872	100	0,965	0,661	0,785	90
Average	0,890	0,910	0,899	0,893	0,872	0,881	98	0,948	0,597	0,701	77

Tableau 1 : comparison between result of content extraction

The following figure allows to visualize better the comparison between the values of F-measure found for each version of the system.

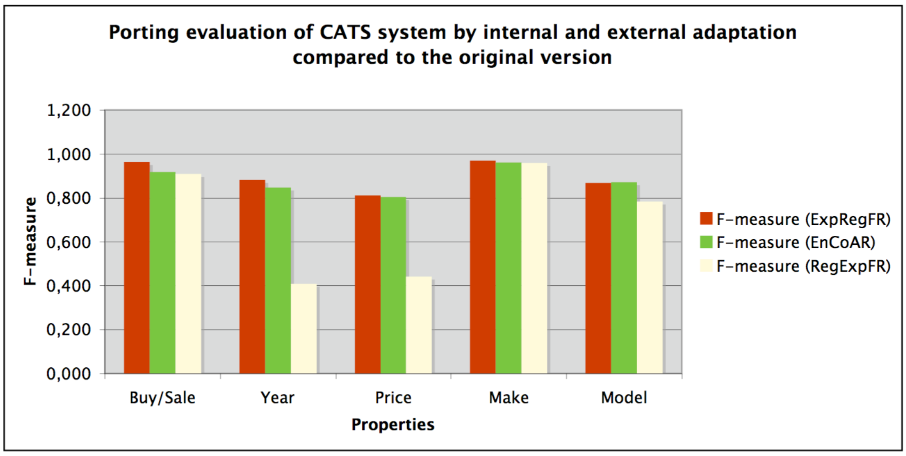


Figure 5 : comparison between F-measure

Conclusion

We chose CATS, as application to be located because it is a platform treating spontaneous SMS in naturel language, which we have access to all its resources. We presented a “external” localization method, which requires a simple access to the internal representation of the application, it gives good results because of the proximity of the sublanguages.

References

Blanchon, H. (2004). Comment définir, mesurer et améliorer la qualité, l'utilisabilité et l'utilité des systèmes de TAO de l'écrit et de l'oral. Une bataille contre le bruit, l'ambiguïté, et le manque de contexte. Grenoble, UJF. HDR: 380. Grenoble, Université Joseph Fourier**:** 380.

Ciravegna, F. (2001). Adaptive information extraction from text by rule induction and generalisation. 17th International Joint Conference on Artificial Intelligence (IJCAI), Seattle.

Daoud, D. M. (2006). It is necessary and possible to build (multilingual) NL-based restricted e-commerce systems with mixed sublanguage and contend-oriented methods. GETA - CLIPS. Grenoble, Université Joseph Fourier**:** 296 pages.

Hajlaoui, N. (2006). Recherche et production de corpus de messages pour la multilinguisation de sites de e-commerce en SMS, initialement en arabe. 6th international IBIMA (International Business Information Association) Conference, Bonn, Allemagne.

Huberman, B., P. Pirolli, et al. (1998). Strong Regularities in World Wide Web Surfing. Science.

Lehman, A. (1996). Construction d'un système de résumé automatique de textes de type scientifique et technique. Récital, Paris.