Methods for porting NL-based restricted e-commerce systems into other languages

Najeh Hajlaoui, Daoud Daoud, Christian Boitet

Methods for porting NL-based restricted e-commerce systems into other languages

[*]Najeh Hajlaoui, [**]Daoud Maher Daoud, [*]Christian Boitet

[*]GETALP, LIG, Université Joseph Fourier
385 rue de la Bibliothèque, BP n° 53
38041 Grenoble, cedex 9, France
Najeh.Hajlaoui@imag.fr,Christian.Boitet@imag.fr

[**]Amman University
PoBox 141009, Zip Code 11814
Amman Jordan
Daoud@ammanu.edu.jo, Daoud.Daoud@imag.fr

Abstract

Multilingualizing systems handling content is an important but difficult problem. As a manifestation of this difficulty, very few multilingual services are available today. The process of multilingualization depends on the translational situation: types and level of possible accesses, available resources, and linguistic competences of participants involved in the multilingualization of an application. Several strategies of multilingualization are then possible (by translation, by internal or external localization etc.). We present a real case of linguistic porting (from Arabic to French) of an e-commerce application deployed in Jordan, using spontaneous SMS in Arabic for buying and selling second-hand cars. Despite the distance between Arabic and French, the localization methods used give good results because of the proximity of the two sublanguages of Arabic and French in this restricted domain.

Introduction

Methods for multilingualizing e-commerce services based on content extraction from spontaneous texts depends on two aspects of the translational situation:

- the level of access 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 no access.
- the linguistic qualification level of the persons involved in the process (level of knowledge of the source language, competence in NLP) and the resources (corpora, dictionaries) available for the new language (s), in particular for the "sublanguages" at hand.

We first discuss the requirement in localizing the Content Extractor (CE, or content extraction module) of such an application, and an analysis of the possible methods. We then present a case study, the linguistic porting (from Arabic into French) of CATS, an application to which we have complete access, so that several multilingualization strategies are possible. In the next section, we present an "external" localization strategy which requires only access to the internal representation. Then, we evaluate this method by comparing it with the results produced by the original monolingual system. We also compare it with the "internal" localization method, which consists in adapting the existing content extractor, and has been described in another paper (Hajlaoui 2007).

1 Multilingualizing NLP-based services

We concentrate on NLP-based systems that perform specific tasks in restricted domains, Figure 1 shows the general structure of these systems. Examples of such applications and services are: categorization of various documents such as AFP (Agence France Presse) reports or customer messages on a SAS (Service After Sale) server, and information extraction to feed or consult a database (e.g. small ads, FAQ, automated hotlines).

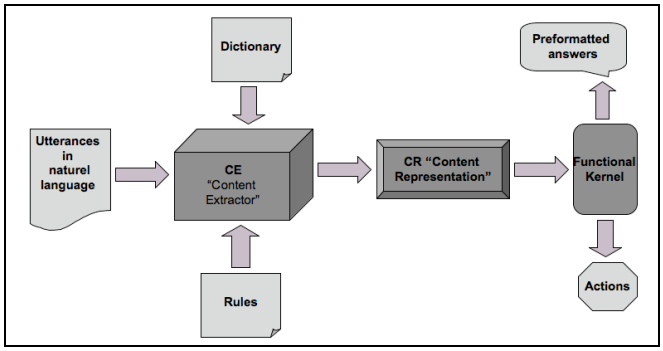


Figure 1: general structure of NLP-based systems

1.1 Benefits and necessity of NL interfaces

The main benefits of building systems based on the processing of spontaneous unedited text are:

- the naturalness of interaction,
- the ability of expressing complex expressions,
- the possibility (more recently) of designing and building domain-focused services based on a domain-specific thesaurus or ontology.

Integrating a module processing small spontaneous task-related texts (ads, messages…) seems to be the only answer to the growing ergonomic needs, especially for e-commerce applications.

The study of the current scene shows that the deployed or operational e-commerce NL interface systems are rare and most of them are only prototypes.
Furthermore, when we find an application, we get very few information about its internal procedure and how its multilinguization has been done or envisaged.
Among the few we found are Pertinence Summarizer (Lehman 1996), a system of automatic summarization 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 CATS "Classifieds Ads Transaction System" (Daoud 2006), a

Arabic-based SMS system for handling classified ads related to buying and selling cars and real estate.

1.2 Sublanguage and content extraction

The reasons why handling spontaneous text is difficult are similar to those encountered in speech processing: non standard grammar (more or less oral), errors (spelling, typing errors), use of some typographical conventions (SMS-specific abbreviations, smileys). Often enough, classified ads or alarms/warnings (road traffic, natural disasters) form a sublanguage which is relatively far from the general language.
The consequence is that it is not possible to use the "analysis approach" and associated tools, developed for written general and clean language, even by specializing them. Rather, it seems necessary to adopt a content extraction approach, using the particularities of the sublanguage at hand.

2 Possible approaches to "porting"

2.1 Common need: a corpus in L2

The primary data needed for porting an application treating spontaneous texts in language L1 to language L2 is a representative corpus in L2 of the same type of data… but that is almost never available. It is hence generally necessary to create one by translation, or by simulation or imagination, and post-edition.
For example, the *Real Estate* part of CATS treats SMS concerning Amman. To *localize* it to French, one should handle cities in France, Belgium, Switzerland, Canada, or Africa etc., and hence one should develop a corpus covering the various corresponding sublanguages. That would be a too expensive process. Because of that problem, we limit our ambition to *porting* and not localizing applications. In the case of CATS, porting to French means that French users could use CATS for sending French SMS in the *same* situation (real estate in Amman), so that a suitable corpus could then be created by translating and post-editing the corpus of available SMS from Arabic to French.
The "good size" of the initial L2 corpus depends on the strategy used. If we adapt an existing CE, the translation of the L1 corpus used for initial development or for regular testing should suffice. If we develop an L2-L1 MT system, we might need a much larger corpus. However, a recent experiment seems to indicate that a SMT usable by the CE to work well can be developed with a rather small corpus, and a complete dictionary.

2.2 Translation

If there is no access to the code, dictionary, and internal content representation of the target application, the only possible approach to localize it from L1 to L2 is to develop an MT system to automatically translate its (spontaneous) inputs from L2 into L1.
This approach might look applicable and straightforward. However, from the practical perspective it is not. As a matter of fact, for restricted domains where we have special sublanguage with specific terminology, traditional MT is not useful.

2.3 Porting or adapting a CE

Each considered application `App` uses a specific CRL (content representation language), say `CRL_app`. Several types of content representation are used, such as property lists (<attribute, value> pairs), typed features structures, logical expressions (Prolog), logico-functional expressions, objects (classes (methods, attributes), instances). Deriving a CE from one application for another one is difficult, even in the same language, because one must guarantee a minimum level of quality, correctness and completeness of the extracted content, as well as the relevance and linguistic adequacy of the produced answers.
Localization at the level of content extraction can be achieved by "internal" porting or "external" adaptation.

2.3.1 Internal CE localization

The first possibility consists in adapting the CE of the application from L1 to L2; but that is viable only if

- the developers agree to open their code and tools,
- the code and tools are relatively easy to understand,
- the resources are not too heavy to create (in particular the dictionary).

That method requires of course training the localization team with the tools and methods used.
Under these conditions, adaptation can be done at a very reasonable cost, and further maintenance (to "follow" the drift of the sublanguage) can later be done cheaply.

2.3.2 External CE localization

The second solution consists in adapting an available CE for L2 to the sublanguage at hand, and to translate its results into the target CRL_app.
For a company wanting to offer multilingualization services, it would indeed be an ideal situation to have a generic CE, and to adapt it to each situation (language, sublanguage, domain, CRL, task, other constraints). However, there are still no known generic CEs of that power, and not even generic CEs for particular languages, so that this approach cannot be considered at present.
A third approach is then to adapt an existing content extractor, developed for L2 and a different domain/task, or for another language and the same domain/task.
We have previously experimented the first method (Hajlaoui 2007) by porting the *Cats* part of CATS from Arabic to French: for that, we adapted its native Arabic CE, written in EnCo, by translating its dictionary, and modifying a few analysis rules.
We also tried the third method, on exactly the same task, and report on that experiment in the following section.

3 Case study: external CE adaptation

3.1 Presentation of CATS

CATS (Classifieds Ads Transaction System) is a platform for buying and selling goods (cars, real estate…) based on the use of Arabic SMS and created by the second author (Daoud 2006). It is deployed by Fastlink (the largest mobile operator in Jordan). It is a C2C based e-commerce system that uses content extraction technology based on sublanguage analysis and knowledge representation to enable SMS users to post and search for classified ads in Arabic. It has two main functionalities: the submission for selling items and the

answering of users' queries through interaction in spontaneous natural language. The system receives an entry in full text without a pre-specified layout, recognizes the various relevant entries, and produces a knowledge representation for further processing. We have two types of users' requests:

- "Sell" post: in which the user is a potential seller.
- "Looking for" post: in which the user is a potential buyer.

Table 1 shows some examples of the car domain, for which we are interested in a first time.

مطلوب سياره هونداي موديل 97 والسعر مابين 3500 الى 3750	Looking for Honda, model 97, price between 3500 and 3750
مطلوب سياره سبور	Looking for sport car
اريد سياره مرسيدس موديل 82 لون ابيض	I want Mercedes car model 82 white color
سياره اوبل فكترا للبيع موديل 2003 فل اوبشن	Opel Vectra car for sale year2000 full option
للبيع سيارة بي ام دبليو 520 لون زيتي فحص كامل م 89 فل عدا الفتحه مرخصه بحال ممتازه بسعر 8500	For sale BMW 520 color dark green full check year 89 full except sunroof licensed in a good condition with a price 8500.
أوبل أسترا ستيشن لون أحمر (بورفتحه سنترزجاج ومريات كهرباء) فحص للبيع .	Opel Astra station color red (power sunroof Center Electrical windows and mirrors check for sale
عندي سياره لاند روفر بدي ابيعها.	I have a Land Rover car I want to sell it
مطلوب سيارة بيجو406	Wanted a Peugeot 406 car
بحاجه لسياره لا تزيد عن 2000 دينار بحاله جيده واقتصاديه في البنزين	In need for a car not more than 2000 dinar in good condition and economical in fuel.
شراء سيارة.	Buying a car
اريدبيع سياره دايو ليمنز موديل 92 فحص كامل فل اوبشن	I want to sell a Daewoo Lemens car year 92 full check full option

Table 1: examples in the cars domain

The overall structure of the CATS reflects both the corpus analysis and the adopted knowledge representation. The CATS system consists of a content extraction (CE) component and a query manager QM component.
The CE component receives SMS text and decodes it into the corresponding knowledge representation CRL-CATS using a domain-specific lexicon. The system is able to extract knowledge from both types of messages.
The QM component takes the KR and converts it into SQL statements. It also issues the SQL statements (query or insert), and checks, validates and formats the results. It also handles situations where no answer found.
One important aspect of this design is that both questions and postings (documents) are processed by the same engine, using the same knowledge representation, leading to accurate matching of questions with answers.
CE is written in EnCo (Uchida and Zhu 1999) and uses a lexicon specific to the domain and a grammar specific to the sublangage.

```
 2000 ريتو ميجان م ;للبيع

;Selling Renault Megane m 2000

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

Figure 2: a CRL-CATS representation

In CRL-CATS (Content Representation Language for CATS), a posted SMS is represented as a set of binary relations between objects. It is a kind of semantic graph with a UNL-like syntax. There are no variables, but the dictionary is used as a type lattice allowing specialization and generalization.

In the preceding example, there is one object, a car (`saloon`), with 4 properties. The first is `sal` (type of post, selling or buying, here `sale` for selling). The other properties are `mak` (make), with value `RENAULT (country<France,country<europe)`, `mod` (model), with value `Megane(country<France, country< europe,make<RENAULT)`, and `yea` (year), with value `2000`.

3.2 Necessity of an "initial" corpus

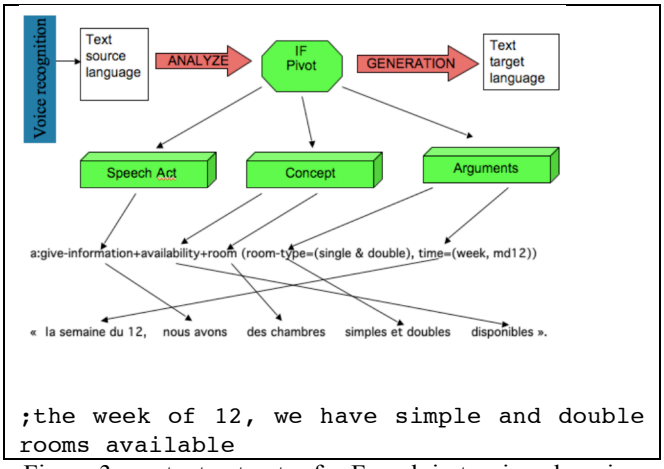
For all localization methods of the CATS system into French, the first thing to do was to collect or build a "French starting corpus", similar to that used by Daoud at the beginning of his project for Arabic. That was obviously necessary to study the syntactic form of the SMS to be treated in French, and to see also which lexical categories to expect. Initially, we used an Arabic corpus generated by CATS and translated it into "French spontaneous SMS", expected 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/literal and natural/functional translations by calculating the edit distance between two translations. The average distance is 21,88 (Hajlaoui 2006), for an average length of 66 characters.

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 combinations of properties and values (make, model, year, colour, 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

We adapted an existing French extractor of 31,918 lines written in Tcl/Tk by H.Blanchon for the Nespole! project (Blanchon 2004), and producing IF (Interchange Format) representations. The IF is a semantico-pragmatic pivot used for restricted domains. The second step was then to translate the IF expressions into CRL-cats graphs.

The IF is a semantico-pragmatic pivot used for restricted domains. Figure 3 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 electronic version of the IF specification (version of 08-18-2002) and the CRL-CATS specification.




```
;the week of 12, we have simple and double rooms available
```

Figure 3: content extractor for French in tourism domain (Blanchon 2004)

3.3.1 Content extraction method in Nespole!

Blanchon's CE uses a method based on finite-state transducers. Relevant sequences are described by regular expressions and attached actions incrementally consume the input and produce an IF expression.

We tried to understand and use the method used in the second module of Blanchon's CE. As Figure 4 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.

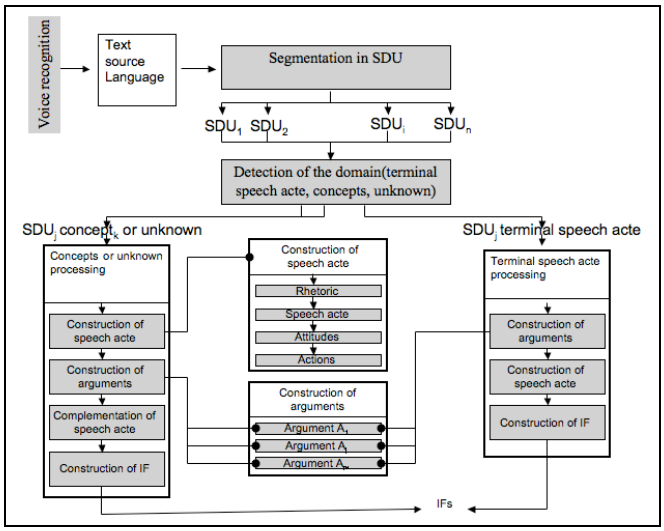


Figure 4: structure of French analysis module into IF for the NESPOLE second demonstrator (Blanchon 2004)

3.3.2 IF and French-IF CE adaptation

We adapted the IF specification to the *Cars* domain. We also enriched it by adding new arguments like vehicle-motor-type, vehicle-hand, etc.

We added new actions, essentially the buying action `e-buy` and the 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.

Most of the work was done on the arguments instantiation stage related to the `vehicle` domain: we instantiate the vehicle specification `vehicle-spec`, as well as other less interesting arguments such as: `theDistance`, `theLocation`, `theDuration`, `theDestination`, `theTime`, `thePrice`…

A new `VehicleSpec2If` function allows search and construction of the arguments related to the `vehicle` concept. The only argument already programmed in Blanchon's CE was `frenchvehicle`, which can have values `voiture`, `ski`, `camion`, `bus`, `train`, `avion`… Likewise, the `Argument2if` function builds the IF associated values. Figure 5 is an example of result obtained after adaptation.

```
Input 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 sale a big French car BM 325 4 doors diesel blue VGS first hand insurance completed with CT without air-cond VG price last mod
```

```
Output1 = {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,     age-vehicle=new_mod,
  vehicle-assurance=insured,
  vehicle-controle=total_check,
  vehicle-air-condition=no_air_condition,
  vehicle-nationality=french,
  price-vehicle=good_price))}
```

Figure 5: result of content extraction on a French SMS

We call the obtained result "IF-CATS" (output1 in the example). We built a compiler, which analyses and transforms the IF-CATS expressions in CRL-CATS graphs by using an "IF-CRL" dictionary which facilitates the substitution of the arguments.

3.3.3 Compiler IF-CATS_CRL-CATS

We built a compiler, which analyses and transforms the IF-CATS expressions in CRL-CATS graphs by using an "IF-CRL" dictionary which facilitates the substitution of the arguments.
Figure 6 gives the same result of the CRL-CATS protected by the compiler. It is the same result as that of Figure 2. 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.

```
 2000 للبيع  رينو ميجان م;

;A vendre RENAULT Megane m 2000
;Selling Renault Megane m 2000

 *********IF-CATS**********
;{a:give-information+concept(
action=e_sell,
vehicle-spec=(car, vehicle-make=RENAULT,
 vehicle-model= Megane, vehicle-age=2000,
vehicle-price=,
vehicle-color=,
vehicle-condition=, vehicle-assurance=,
vehicle-controle=, vehicle-air-condition=,
vehicle-size=,
vehicle-motor-type=,
vehicle-hand=,
vehicle-nationality=, vehicle-mileage=))}

 *********CRL-CATS**********
S
sal(saloon,sale)
mak(saloon,RENAULT(country>France,country>
europe))
mod(saloon,Megane (country>France,country>
europe,make>RENAULT))
yea(saloon,2000)
/S
```

Figure 6: an IF-CATS_CRL-CATS compiler result

3.3.4 Results and evaluation

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 translated manually the corpus used for the evaluation of CATS Arabic version (original). We computed the recall R, the precision P and the F-measure F for each most important property (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)

Table 2 summarizes the results obtained and Table 3 shows details (Hajlaoui and Boitet 2007). Properties having numbers as values, like price and year, lower the percentage of porting by external adaptation, but the advantage is that that method requires only to access the internal representation of the application.

Porting by	minimum	average	maximum
internal adaptation	95%	98%	100%
external adaptation	46%	77%	99%
statistical translation	85%	93%	98%

Table 2: evaluation of three localization methods used for porting CATS_Cars from Arabic to French

Figure 7 allows to better visualize the comparison between the values of F-measure found for each version of the system.

[1] EnCo is a tool based on rules and dictionaries used for content extraction in original version of CATS system.
[2] Standard page = 250 words

	EnCoAR (original version)			EnCoFR (internal adapatation)				RegExpFR (external adapatation)				SMTFR (adapatation by translation)			
Properties	Precision	Recall	F-measure (EnCoAR)	Precision	Recall	F-measure (EnCoFR)	% porting	Precision	Recall	F-measure (RegExpFR)	% porting	Precision	Recall	F-measure (SMTFR)	% porting
Buy/Sale	1,0	1,0	1,0	1,0	0,9	0,9	95	1,0	0,8	0,9	95	1,0	0,8	0,9	92
Year	0,8	1,0	0,9	0,9	0,8	0,8	96	0,8	0,3	0,4	46	0,8	0,7	0,7	85
Price	0,8	0,8	0,8	0,8	0,8	0,8	99	1,0	0,3	0,4	55	0,9	0,7	0,8	98
Make	1,0	1,0	1,0	1,0	1,0	1,0	99	1,0	0,9	1,0	99	1,0	0,9	0,9	96
Model	0,9	0,8	0,9	0,8	0,9	0,9	100	1,0	0,7	0,8	90	1,0	0,7	0,8	95
Average	0,9	0,9	0,9	0,9	0,9	0,9	98	0,9	0,6	0,7	77	0,9	0,8	0,8	93

Table 3: comparison between result of content extraction

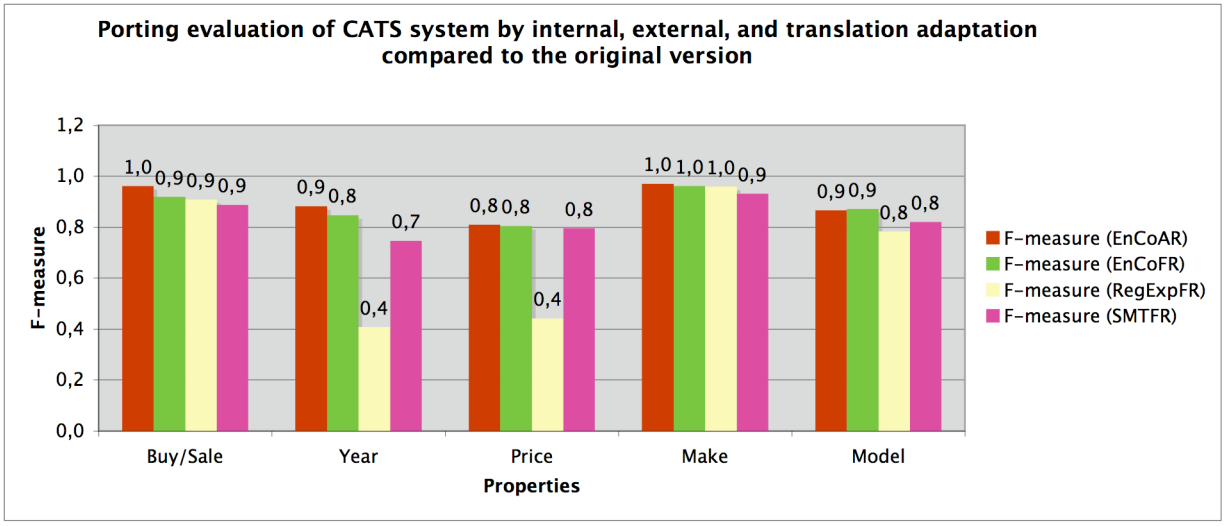


Figure 7: comparison between F-measures

Conclusion

We have presented several possible methods for "porting" applications based on handling the content of spontaneous NL messages in a "native" language L1 into another language, L2. In a previous paper, we described an experiment and an evaluation of the "internal" strategy, consisting in adapting the native CE (content extractor) to L2.

Here, we reported on another case study, in which we experimented the "external" strategy, consisting in adapting an existing CE for L2 to the domain and to the CRL (content representation language) of the target application. The evaluation was done similarly, on the same part of the CATS system, *Cars,* to port it from Arabic to French. Porting by translating messages from L2 to L1 (here from French to Arabic) has also been done, and its evaluation is also given.

An interesting conclusion of these experiments is that, in a real case of linguistic porting, all three localization methods used gave good results. The most likely reason for that seems to be that, although Arabic and English are quite distant as "general languages", their considered sublanguages are quite similar. That corroborates the analysis made by (Kittredge 1986).

Acknowledgments

We would like to thank our reviewers, who made many constructive remarks and suggestions, which we gladly incorporated in this new version of the paper.

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, Université Joseph Fourier. Thèse de HDR: 380 p.

Ciravegna, F. (2001). Adaptive information extraction from text by rule induction and generalisation. IJCAI, Seattle.

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

Hajlaoui, N. (2006). Recherche et production de corpus de messages pour la multilinguisation de sites de e-commerce en SMS, initialement en arabe. proc 6th international IBIMA Conference, Bonn, Allemagne 11 p.

Hajlaoui, N. and C. Boitet (2007). Portage linguistique d'applications de gestion de contenu. TOTh Conférence sur la Terminologie & Ontologie: Théories et Applications, Annecy France 13p.

Hajlaoui, N. (2007). Multilinguïsation de services de gestion de contenu. IC "Ingénierie des Connaissances", plate-forme AFIA "Association Française pour l'Intelligence Artificielle", Grenoble, France 2p.

Huberman, B., P. Pirolli, et al. (1998). Strong Regularities in World Wide Web Surfing. Science vol 280, pp 95-97.

Kittredge R. (1986) Analyzing Language in Restricted Domains. In "Sublanguage Description and Processing", R. Grishman & R. Kittredge, ed., Lawrence Erlbaum, Hillsdale, New-Jersey, 248 p.

Lehman, A. (1996). Construction d'un système de résumé automatique de textes de type scientifique et technique. RECITAL , Paris pp 65-69.

Uchida, H. and M. Zhu (1999). Enconverter Specifications, UNU/IAS UNL Center, 33 p.