Experiments on Building Language Resources for Multi-Modal Dialogue Systems

Laurent Romary*, Amalia Todirascu[§], David Langlois*

*INRIA Lorraine, Batiment LORIA, Campus scientifique, 54506 Vandoeuvre-les-Nancy cedex, France,

{Laurent.Romary,David.Langlois}@loria.fr

§ Université de Technologie de Troyes, 12, rue Marie Curie, 10010 Troyes cedex, France, Amalia.Todirascu@utt.fr

Abstract

The paper presents the experiments made to adapt and to synchronise the linguistic resources of the French language processing modules integrated in the MIAMM prototype, designed to handle multi-modal human-machine interactions. These experiments allowed us to identify a methodology for adapting multilingual resources for a dialogue system. In the paper, we describe the iterative joint process used to build linguistic resources for the two cooperative modules: speech recognition for speech modality and syntactic/semantic parsing.

1. Introduction

This paper focuses on the identification of a methodology for adapting linguistic resources for a human-machine dialogue system. The prototype resulting from the European Multimedia Information Access using Multiple Modalities (MIAMM) project ((Reithinger and all, 2003), (Kumar and Romary, 2002)) proposes to the human user several modalities to explore a music database: speech, haptic interfaces, visualisation, combined into a human-machine dialogue system.

Such human-dialogue system requires a language model designed for the application. While the MIAMM project integrates innovative haptic modules, we have been confronted to the lack of real user-system interactions. It is difficult to find annotated dialogue corpora for a specific domain (containing only speech and text), multimodal corpora including haptic interaction are not yet available. Building annotated dialogue corpora is very expensive (Rapp & Strube, 2002). Due to the fact that the haptic interfaces were not available at the beginning of the project, we had to develop suitable linguistic resources.

We present here the adaptation process of our tools and pre-existing linguistic resources for this project to provide language models for the speech modality (speech recognition) and for the parser (used to build a semantic representation of the speech input).

Across the various languages (French, English and German) used in the MIAMM project, we tried to maintain the same linguistic coverage, even if the actual implementations of various parsers and speech recognisers were different. For this purpose, the speech recognisers and the parsers use a shared language model (a shared vocabulary and grammar), established on the basis of user scenarios.

We tested several methods for speech and text processing. We use two robust parsing methods for information extraction: template-based parsing and TAG-based grammars. For English and German, we use the same speech recogniser, together with the SPIN template-based parser. For French, we use the ESPERE speech recogniser (Fohr and all, 2000) and a LTAG parser (Lopez's parser (Lopez, 1999)) using local grammars in order to extract the semantic interpretation. The speech recognisers output wordgraphs containing most probable sentences (in MPEG-7 format), the SPIN parsers process them and provide semantic interpretations to the Dialogue Processing Manager. All these modules use a shared language model and a similar linguistic coverage.

This paper illustrates the work done for the French modules, even if the actual prototype includes English, German and French languages. We chose the French modules in order to illustrate the adaptation process of modules implementing different approaches: statistical methods for speech recogniser and classical linguistic processing approach based on TAG grammars (Joshi, 1987) for parsing.

2. The French modules

We present the main features of the French modules interpreting speech input and providing a semantic interpretation according to the domain model.

The ESPERE speech recognition system is used for acquiring/recognising vocal commands from user. Its output is a word lattice (in MPEG7 format) containing the n-best possible sentences matching the acoustic input. ESPERE relies on the HMM technology (Kai-Fu & Fileno, 1992) and is dedicated to small vocabulary applications. Basically, the system is made up of two modules: (1) the acoustic module is composed of 40 monophones trained on the BREF80 database (Lamel and all, 1991); (2) the language model is a statistical bigram model (Jelinek, 1990), but more performant language models can be used for parsing the word lattice (as it is done in the MIAMM project).

The Lopez parser (Lopez, 1999), used for interpreting the output of the speech recogniser, is based on the Lexicalised Tree Adjoining Grammar (Joshi, 1987) formalism. We chose this parser because it provides partial parsing results (in order to handle noisy or erroneous input) and because LTAG represent words in their syntactic context (helping us to build a semantic interpretation). The parser use general French grammar validated by linguists, described in Tree Adjoining Grammar Markup Language (TAGML) format (Pardo and all, 2000). Using the information provided by syntax, we added links to the MIAMM's domain-specific ontology, for obtaining a relevant semantic interpretation, in MMIL format (Kumar and Romary, 2002).

3. Creating/adapting linguistic resources

The methodology used for adapting/creating the language models for our project follows the steps presented in Figure 1. To build the language resources, we stemmed on basic interaction scenarios, while lacking real interaction

corpora. We concertated the efforts of building the grammar and the vocabulary to have similar coverage across languages. For each language, one group designed a context-free grammar to cover these scenarios. The technical vocabularies were extracted from scenarios and grammars and were translated for each language to maintain the same semantic coverage. The statistical language model used for speech recognition was developed directly from these resources. The LTAG parser's resources (used to build semantic representations of speech input) were developed from general LTAG resources and adapted to the application by comparing the linguistic coverage with the French context-free grammar.

While scenarios changed several times during the project, we used an iterative joint process to update resources and language models in order to match the application requirements.
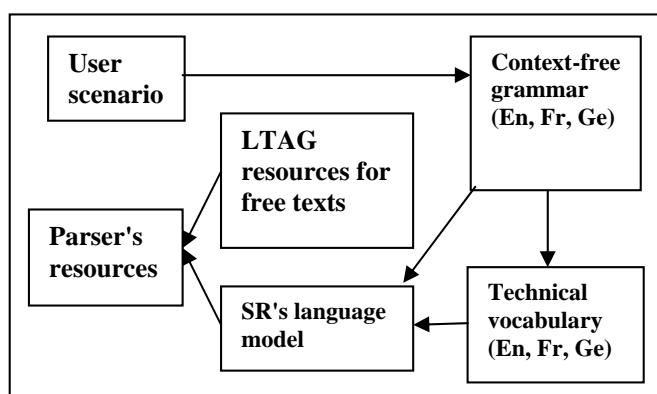

Figure 1. Adapting MIAMM linguistic resources

3.1. User scenarios

Due to the various languages and NLP techniques integrated in the system, we wanted to build language resources covering the same linguistic phenomena in all the languages. We preferred to have an uniform linguistic coverage instead of having only a semantic coverage, as most of the dialogue systems propose (Rapp & Strube, 2002). A homogeneous linguistic coverage consists of several styles (or registers - familiar, elaborated), specific phrases (politeness phrases, time intervals - "from the sixties"), various syntactic components (passive constructions, relative clauses, questions and ellipses) as well as dates or names. We treat identically similar linguistic phenomena in every language. This method assures that the semantic coverage is also similar. The most difficult task was to identify the most significant linguistic phenomena to be handled by the language modules (Wilks and all, 2000).

Due to the lack of some functionalities (haptic and visual interaction), the MIAMM's human factor team proposed possible user scenarios. The scenarios contain possible user interactions involving one or several modalities: haptic, graphic or speech. We do not have real data for training the system, so we replace it with made-up training data.

From the initial scenarios, we identified the syntactic elements and the required vocabulary: some basic predicates, domain-specific objects (database's specific categories: songs, titles, styles, albums etc.), auxiliary phrases (opening session items, closing phrases, referential mechanisms - alterity, similarity, politeness expressions), modality specific vocabulary (visualisation styles, visualisation predicates etc.).

The advantage of this user scenario-based approach is that each developer adapts the resources independently and he decides himself which new entries to be added to the existing lexicons and grammars. The parsers and the speech recognisers could be tested independently for each language, without waiting the other teams. The drawbacks of this approach are the requirements of building exhaustive user scenarios (impossible while some functionalities are not available yet), as well as the different stages of development of the various modules.

3.2. Designing the language models

3.2.1. Creating a training corpus for the speech modality
The bootstrap of a bigram model, used by the speech recogniser, is a training corpus relevant to the task. Unfortunately, as explained in the introduction, such a corpus was not available and we had to remedy this lack.

In order to generate a training corpus, we designed a context-free grammar. By developing this grammar, our objective was to benefit from the compactness, the flexibility of this formalism to model a language allowing a wide range of possibilities for user to utter commands and requests. This grammar contains almost 200 rules and is based on a 400 word vocabulary.

3.2.2 Training the bigram model
For training the language model, it was not possible to collect the bigram frequencies directly from the corpus generated with the grammar, because this corpus was too huge. Rather, we partially generated the training corpus at a class level. These classes were chosen among non-terminals. For example, one sentence of this training corpus is:

"donne-moi le GENRE des années DECADES" (give me the GENRE of the DECADES's)

With this corpus, we assumed a uniform distribution of the words into each class. For example:

$$P\left(90 \mid ann\acute{e}es\right) = \frac{P\left(DECADES \mid ann\acute{e}es\right)}{\mid DECADES \mid}$$

In the following sections, we describe several methods to estimate the bigram probabilities and give the performance of the speech recogniser for each method.

3.2.3 Adapting the TAG parser
Lopez's parser has an initial domain-independent lexicon and grammar, not very useful in the context of multi-modal musical search. We add domain-specific words or words designing several types of searches in the musical database (by similarity, by musical dimensions: mood, style, genre) to the lexicon, and new domain-specific lexical categories (used to build specific syntactic components: a style followed by a mood and by a time interval, a request verb followed by a similarity search). We added new lexical entries specific to various human-machine interactions (haptic, visual).

The parser's output (derivation trees and derived trees) are used to build a semantic representation in MMIL format (Kumar and Romary, 2002). MMIL elements contain several events and participants and relations between these elements. The relations correspond to the syntactic structure represented by each elementary tree. A mapping between the various lexical entries and the domain-specific ontology was required to build the appropriate semantic representation. We inspected the context-free grammar's rules and we generated specific local grammars (elementary trees tagged with semantic relations, by using a meta-grammar (Gaiffe and all, 2003)), for modelling each specific phenomena. MMIL specifications changed also during the project, so several elementary trees have been added (alternatives, time periods); some morphological features (mode, tense) have been modified in order to handle the changes.

The main changes of the grammar concern the preference for using substitutions instead of adjunctions in order to reduce the number of parsing results. The use of substitutions reduces the number of possible parsing results, in order to increase parser's efficiency. If a substituted syntactic component missed, it is interpreted as an empty MMIL participant or event.

The linguistic coverage concerns several possible combinations of the following syntactic components: elliptic phrases (*celui-ci*, *celui-là*), domain-specific noun groups (*du* GENRE, *du* GENRE MOOD, *une liste de chansons/albums*, TITRE, ARTIST), opening and closing events (*commence*, *annule*, *oui*, *non*), demand verbs (*demander*, *vouloir*), navigation verbs (*avancer*, *afficher*, *déplacer*, *montrer*), very simple negation (only to cancel the previous orders).

4. Recognition experiments

In this section, we describe several ways to estimate the bigram probabilities and give the performance of the speech recogniser for each of these ways.

As the speech recogniser is integrated into the general architecture of the MIAMM project, the evaluation should be an user-centered evaluation. But such an experimental protocol is not ready for the moment. So we decided to evaluate the system in terms of Word Error Rate. This evaluation is required because speech recognition accuracy must be high to build an effective dialogue with the user. Too many errors at the recognition step are not acceptable.

4.1. Experimental protocol

We recorded 88 sentences that can be parsed by the TAG parser, e. g. that can be generated by the grammar. These sentences were selected to cover the most possible linguistic phenomena. We remark that, even if we decided to give enough liberty to the user for the speech modality, each acceptable phrase will be parsed. Therefore, we decided to not use out-of-application sentences, and out-of-vocabulary words.

The sentences were recorded by 4 speakers, 2 females (OM and AB) and 2 males (KS and DL). Each of them recorded 22 sentences.

4.2 Estimating bigram probabilities and evaluation

In this section, we describe several methods to estimate the bigram probabilities. For each method, we evaluate the corresponding speech recognition system by the Word Error Rate on the 88 sentences. Two parameters are used to integrate the language model into the system: the language model's weight in comparison with the acoustic models; an additional cost added to each bigram in order to prevent from too many insertions. In the following experiments, the results are given for the best values for these parameters.

4.2.1 Estimation 1

The first idea consists in estimating the bigram probabilities by using directly the bigram frequencies from the training corpus. The performances are given in Table 1. We can first remark that the WER is low. A study of the errors shows some confusions between very acoustically closed words (*1980* and *1981*, *veux-tu* and *peux-tu*). Globally, these errors do not modify the overall semantics of the sentences.

WER (speaker, error rare)			
OM, 2.7	KS, 4.8	AB, 3.7	DL, 4.3
Overall error rate : 3.8		**Standard deviation: 0.9**	

Table 1 : performance for Estimation 1

4.2.2 Estimation 2

The second method makes the hypothesis that the probabilities may be not representative of a real life use because the training corpus has been generated from the grammar. In order to check this hypothesis, we evaluated a system where all bigrams have the same probabilities. But, in this model, bigrams which do not occur in the training corpus are given a null probability. So, this model gives only a binary information: a given bigram is part or not of the application's language. The performances are given in Table 2. We remark that WER increases a bit, but the increasing is not significant. This evaluation tends to confirm that real life probabilities may be not important (for this experiment).

| WER (speaker, error rare)			
OM, 3.5	KS, 4.3	AB, 3.7	DL, 4.3
Overall error rate : 4.0		**Standard deviation: 0.4**	

Table 2: performance for Estimation 2

4.2.3 Estimation 3

For the following experiment we abandon the bigram constraints given by the grammar. For that, we used the Good Turing discounting so that all bigrams get a not null probability. The discount is applied to the bigram frequency from the training corpus generated with the grammar. This method is the first step towards a model less dedicated to the application, even if the vocabulary remains the same. The results of the system using this language model are very bad compared to the ones described in this paper. We can conclude that the constraints given by the context-free grammar are necessary, even at bigram level.

4.2.4 Estimation 4

Last, we tried to extract the bigrams probabilities from a general, free language corpus. We extracted the bigrams probabilities for bigrams present both in the general corpus and the application's corpus. As general corpus, we chose 2 years of the French newspaper « Le Monde ». We used a linear combination between the two models (GM for the not specific (General) Model, and AM for the Application's Model). This way is a kind of language model adaptation (Bellagarda 2004). The performances of the linear combination for several values of the AM's weight are given in Figure 2. This figure shows that the bigram probabilities estimated from "Le Monde" lead to worse results when the weight of this corpus increases. This indicates that the general model generalizes too much the syntactic features of the application. The dedicated context-free grammar must be the central bedrock of the language model. Using more general language need specific adaptation processes. One important point is that such process should take into account the necessary homogeneity with the LTAG grammar's language.
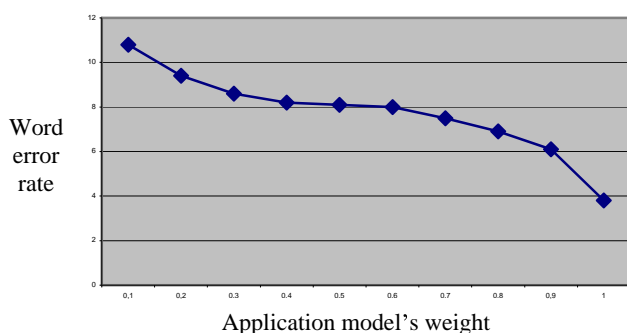
Figure 2: WER for several linear combinations between GM and AM

5. Parsing Experiments

During the iteration phase, we refined the parser's linguistic resources by interacting with other modules. Visualisation and haptics provided new functionalities, so we added new lexical entries, specific lexical categories (VISUALISATION_MODE, DIRECTION) and specific elementary trees (for specific navigation commands, for time intervals)

After testing the parser and the SR, we need to synchronise the language model and the parser's language resources in order to cover the same training corpus. The vocabularies of the two modules are now very similar, after completing them with missing flexed forms or syntagms.

Dialogue Manager module uses a domain ontology to decide which action to do as the answer to the user's requests. Domain ontology changed several times during the project; we had to re-generate the mapping between lexical entries and the domain concepts.

French parser is quite slow compared to the other parsers (for German and for English), due to the fact that the TAG grammar is large (contains a lot of elementary trees for specific phenomena). But, even if partial parsing is provided, the parser builds some MMIL components.

6. Results and further work

The ESPERE speech recogniser and the TAG parser cover the same linguistic phenomena and share the same lexicon, due to the use of shared user scenarios. The relevance of the test corpus will be evaluated by comparing with real user input from the MIAMM prototype, but it helped us to adapt the language modules in the absence of well-defined system's specifications. Further work will focus on the evaluation of methodologies for building test suites, in the context of a multi-modal dialogue system.

The MIAMM project involves our two teams: the "Langue et Dialogue" group which aims at building human-machine dialogue systems, and the Speech Group which aims at speech recognition. This project is the first step towards a collaboration based on the use of formal language/dialogue models during the speech recognition process.

References

Bellagarda J. R. (2004) Statistical language model adaptation : review and perspectives. Speech Communication. 42 (2004) pp. 93-108.

Fohr, D., Mella, O., Antoine, C. (2000). The automatic speech recognition engine ESPERE : experiments on telephone speech. In Proceedings of ICSLP 2000. v. 4, pp. 246-249

Gaiffe B., Crabbe B., Roussanaly A. (2003) Meta-grammar Compiler. In Proceedings of TAG+6, Venice.

Jelinek, F. (1990) Self-organized language modeling for speech recognition. In A. Waibel, K.-F. Lee (Eds.), Readings in Speech Recognition. Pp 450-506.

Joshi A.. (1987). An Introduction to Tree Adjoining Grammars. Mathematics of Language.

Kai-Fu L., Fileno A. (1992). Continuous speech recognition. In S. Furui & M. Mohan Sondhi (Eds.), Advances in Speech Signal Processing. pp 623-650.

Kumar A., Romary L. (2002). A Comprehensive Framework for Multimodal Meaning Representation. In Workshop on Computational Semantics (IWCS-5), Tilburg, Netherlands.

Lamel, L., Gauvain, J.-L., and Eskenazi, M. (1991). BREF, a large vocabulary spoken corpus for french. In European Conference on Speech Communication and Technology, Genova, Italy. pp. 505-508.

Lopez P. (1999). Robust Parsing with Lexicalized Tree Adjoining Grammars. Ph.D.Thesis, INRIA, Nancy, France.

Pardo, M.A., Seddah, D., de la Clergerie, E. (2000). Practical aspects in compiling tabular TAG parsers. In Proceedings of the TAG+5 Workshop, Université Paris 7, France.

Rapp, S, Strube, M. (2002). An Iterative Data Collection Approach for Multimodal Dialogue Systems. In Proceeding of LREC'2002, Canary Island pp. 661-665.

Reithinger N, Fedeler, D., Kumar, A., Lauer, C., Pecourt, E., Romary, L. (2004). MIAMM- A Multi-Modal Dialogue System Using Haptics. In J. Van Kuppevelt, L.Dybkjaer, N.O. Bernsen (Eds.) Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, Kluwer Academic Publisher.

Wilks, Y., Catizone, R., (2000) Human-Computer Conversation. In Encyclopedia of Microcomputers, Dekker, New York.