Interchange Format-Based Language Model For Automatic Speech Recognition In Speech-To-Speech Translation

Minh-Quang Vu, Laurent Besacier, Eric Castelli, Brigitte Bigi, Hervé Blanchon

Abstract—**This paper relates a methodology to include some semantic information early in the statistical language model for Automatic Speech Recognition (ASR). This work is done in the framework of a global speech-to-speech translation project. An Interchange Format (IF) based approach, representing the meaning of phrases independently of languages, is adopted. The methodology consists in introducing semantic information by using a class-based statistical language model for which classes directly correspond to IF entries. With this new Language Model, the ASR module can analyze into IF an important amount of dialogue data: 35% dialogue words; 58% speaker turns. Among these 58% turns directly analyzed, 84% are properly analyzed.**

I. INTRODUCTION

In automatic speech understanding or translation system, the role of Automatic Speech Recognition (ASR) is to obtain a text hypothesis from a speech signal while, generally, this hypothesis is further treated by a separate understanding or analysis module, which transforms the text string into a semantic representation. Both ASR and understanding (or analysis) modules use linguistic resources like dictionaries, language models and/or grammars, but they are often seen as "black-boxes" to each other. Even if some works (see Vermobil [1] or SLT [2]) report a real resource sharing or intelligent interfacing between ASR and analysis, to our knowledge, very few experiments have been carried out to include some semantic information early in the ASR module.

This paper proposes to include some semantic information early in the statistical Language Model (LM) for ASR. This work is realized in the framework of a global speech-to-speech translation project called NESPOLE[1] [3]. Within this project, an Interchange Format (IF) based approach, representing the meaning of phrases independently of languages, was adopted for the actual translation. This pivot-based approach has several advantages and potentialities. The most obvious advantage is the reduction of the number of different systems, which have to be implemented. Given *n* different languages, an analysis chain (starting from the spoken input and delivering an IF representation) and a synthesis chain (taking the IF representation and providing a linguistic form for it) for each language suffice to yield a system capable of dealing with speech-to-speech translation between all of the possible language pairs. That is, the resulting system would require *n* separate analysis and synthesis chains, instead of the otherwise required quadratic number of modules. Furthermore, given that each module involves only one language, native speakers of that language can do the development. Another important advantage concerns portability to a new language; given the described configuration, a lower effort is necessary to make an existing system capable of dealing with a new language. This strikingly contrasts with the case of a direct or transfer-based translation technique. In the first case, the addition of a new language implies the construction of *2n* new complete modules to link, both ways, each old language to the new one. In the second case, the addition of a new language implies the construction of an analysis and generation modules for the new languages and *2n* transfer modules to link both ways each old language to the new one.

The IF [4] relies on dialogue acts, concepts, and arguments. Dialogue acts describe speaker's intention, goal, and need. Concepts define the focus of the dialogue act. Several concepts may appear in one IF. Arguments instantiate discourse variable values. An IF is encoding a Semantic Dialogue Unit (SDU), thus a dialogue turn may have to be described with several IFs. The IF focuses more on the intent rather than the literal meaning of the utterance. For an utterance meaning "I'd like a room that costs 70 euros", the IF would be:

```
c:give-information+disposition+price+room
(disposition=(who=i, desire), price=(quantity=70,
currency=euro), room-spec=(identifiability=yes, room))
```

The global architecture for speech translation using the IF approach is described in Figure 1.

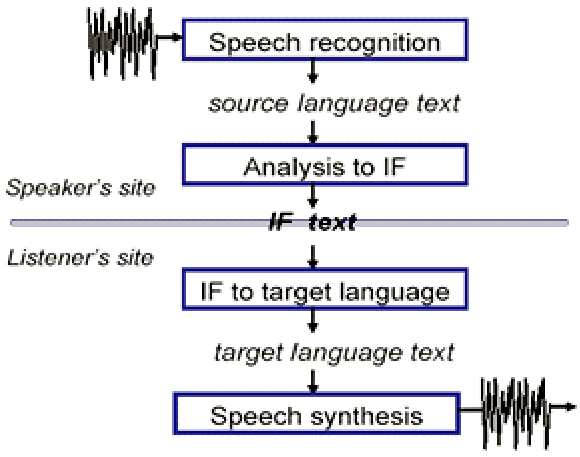

Figure 1: Overall components interaction in global speech to speech translation system

Minh-Quang Vu, Laurent Besacier, Brigitte Bigi, Hervé Blanchon: CLIPS-IMAG Lab. UJF, BP53, 38041 Grenoble cedex 9, France (email: {quang.vu-minh, laurent.besacier, brigitte.bigi,herve.blanchon}@imag.fr}

Eric Castelli: MICA International Research Center. 1 Dai Co Viet - Hanoi, Vietnam (email: eric.castelli@imag.fr)

[1] see http://nespole.itc.it/

This work is at the interface of *speech recognition* and *analysis to IF* modules. More precisely, the speech recognition module was adapted to be able to deliver a chain partially or completely analyzed into IF. This was done by using a class-based statistical language model for which classes directly correspond to IF entries. In the method proposed in this paper, only the most frequent IFs are selected to represent a class in the language model The methodology used to obtain this "semantic" language model is described in section 2 of this paper. Section 3 presents some preliminary experimental results obtained with this methodology while section 4 concludes this work.

II. LANGUAGE MODEL CONSTRUCTION

We now address the problem of predicting a word from previous words in the context of ASR system. Many researches have proved that "classes", generally obtained by clustering, can improve performance of various natural language processing tasks. Clearly, some words are similar to other ones in their meaning and syntactic function. Classes have been used to construct interpolated 3-gram class-based language models. Some examples can be found in [7][8]. Various methods can be used for grouping words together to class according to the statistical similarity of their surrounding. As defined in [9], there're three types of clustering algorithms. The first is a type that uses various heuristic measure of similarity between the element to be clustered and has no interpretation as a probability model. The second type has a clear interpretation as a probability model, but no criteria to determine the number of clusters. The third has interpretation as a probability model and uses some statistically motivated model selection, criteria to determine the proper number of clusters. In our case, the ASR is only a module of a global speech-to-speech translation system. This implies the class-based language model construction fall in none of them. It is different in both construction methodology and obtained results. By introducing the use of "IF classes", there will be two advantages:

- like other class-based language model, it improves the ASR results
- classes are deducted automatically from IF-analyzed corpus
- Output sentences will be partially analyzed into IF, which mean less time-consuming of the global process.

We present now how to construct the semantic language model. This involves two steps: (1) the selection of the most frequent IF classes to be included in the LM and (2) the LM calculation itself. The following details these two steps which are all full automatic.

A. Most Frequent If Classes Selection

Frequent-IF classes means IF components which appear the most frequently during a dialogue. This is because in a dialogue, there are some semantic units which are repeated more frequently than some others, and there are many semantic units which appear in almost every dialogue. In this step, we find out these frequent IFs and regroup them into classes corresponding to IF semantic entries (example of these classes are dialog acts like *acknowledge*, *affirm*, *negate* …). Figure 2 shows how these frequent IF components selection was achieved automatically.

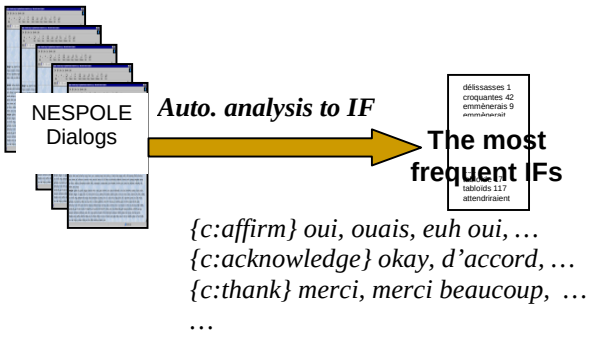

Figure 2: Most frequent IFs selection

A robust, pattern-based, automatic IF analyzer [5] is used to automatically analyze a development corpus made up of 46 dialogue transcriptions collected during the NESPOLE project [2]. This corpus contains many possible dialogues between a client and a travel agency about prices, hotel and ticket reservation for instance. The automatic analysis to IF transforms all of these dialogues to IF language representation. We then have a French-IF aligned corpus. It is however not perfect since automatic analysis obviously makes errors, but we make the hypothesis that despite these errors the selection of the most frequent IF components is correct. Then, we regroup the aligned data by IF and list all SDUs corresponding to a same IF, obtaining the semantic classes as shown in table 1. For instance, the *affirm* class will contain different variants representing a same meaning in French.

The number of semantic classes obtained by this automatic process is important, but taking into account the frequency of occurrence and the size (i.e. the number of variants for a same class) of these classes, only 41 classes are finally retained.

IFs CLASSES	Example SDU	Percentage in total 3194 SDUs
{c:affirm}	Oui, ouais, mouais…	22%
{c:acknowledge}	d'accord, entendu, ok…	19%
{c:exclamation (exclamation=oh)}	Oh, ah, ha….	4%
…	…	…

Table 1: Examples of frequent IF-classes

B. Language Model Estimation

Having obtained the list of semantic classes corresponding to frequent IF entries as illustrated in figure 2, we use it in combination with the language model training data to build our new LM as shown in figure 3.

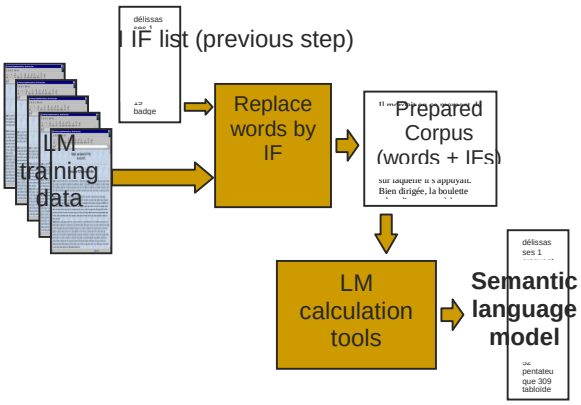

Figure 3: IF class-based language model learning method

On the language model training data which composes of 46 NESPOLE dialogues, we replace all words which are elements of our new semantic classes by the name of the class itself. It means replace french semantic units by their IF representation equivalents. This result in a "prepared" training corpus which contains both french words and IF-language entries. Next, the traditional trigram language model calculation tools are used and give our new language model.

Having the new LM properly built, we have integrated and tested it in the complete speech recognition system. The experimental results are presented in the following section.

III. EXPERIMENTAL RESULTS

A. *ASR System Description*

Our continuous French speech recognition system RAPHAEL uses Janus-III toolkit [6] from CMU. The context dependent acoustic model was learned on a corpus that contains 12 hours of continuous speech of 72 speakers. The vocabulary contains nearly 1500 lexical forms: some lexical forms are specific to the reservation of the tourist information domains whereas the other words are the most frequent words that can be encountered in the French language. More details on the French ASR used in NESPOLE can be found in [3]. The LM of this system uses classes selected manually; moreover, elements of these classes were selected manually, too.

For contrastive tests, two language models learned on the same training corpus were calculated: one using classes constructed manually and one with semantic classes, obtained with the automatic methodology described in section 2. We've compared after that these two results for evaluation.

B. *Test Corpus*

The test signals are 216 speaker turns extracted from the dialogue corpus collected during the NESPOLE project. Table 2 shows examples of these test speaker turns (they are initially in French; here we translate them into English for readers not familiar with French). We also show in the second column the hypothesis strings obtained as output of the recognition module with our semantic language model. We see that some simple speaker turns were analyzed completely to IF. There are also some others more complex speaker turns which were partially or totally analyzed to IF.

Reference sentences	ASR output with our new LM
oui je vous entends[2]	c:affirm c:dialog-hear(who=i, to-whom=you)
euh je vous entends pas très fort mais c'est correct[3]	euh c:dialog-hear(who=i, to-whom=you) pas très forme ce_qu on est
oh oui c'est bon[4]	c:exclamation (exclamation=wow) c:affirm c:acknowledge
oui[5]	c:affirm
d'accord[6]	c:acknowledge

Table 2: Examples of hypothesis strings obtained as output of the ASR module with our new LM

C. *ANALYSIS OF THE RESULTS*

1) Error Rate Comparison

First, in order to verify that these changes in the recognition module allowing a partial analysis to IF do not degrade the performance of the initial speech recognition system, we have compared the error rate between initial system and our new system. The Word Error Rate (WER) obtained with the initial system using classes constructed manually was 31.9% while the WER obtained using the new LM, and after reconstructing French text from the IF-classes, is 32.9%. Thus we can say that the new LM does not introduce significant ASR performance degradation.

2) Statistics On Early If Analysis During ASR

The 216 test speaker turns were made up of 915 words. Among these 915 words, 35% were directly analyzed into IF early in the ASR module.

If we look at the speaker turns, 125 turns among 216 (58%) were directly analyzed into IF early in the ASR module. Of course, these are mainly the shortest dialogue turns which were totally analyzed during ASR, but these results are encouraging since a significative part of the IF analyzer may be saved by the ASR module using IF class-based language model.

If we look more precisely to the data, among these 58% turns directly analyzed, we find 84% turns properly analyzed. This percentage corresponds to the ASR sentence error rate on our test set.

IV. CONCLUSION

We have presented a new methodology for automatically introducing some IF-classes into a statistical language

[2] yes i can hear you
[3] hum i can not hear you very well but it's okay
[4] oh yes that's fine
[5] yes
[6] okay

model. This “semantic language model” was tested in the framework of a speech-to-speech translation project. With our new IF class-based language model, the ASR module can analyze into IF an important amount of dialogue data: 35% dialogue words; 58% speaker turns. Among these 58% turns directly analyzed, 84% are properly analyzed. In future works, the analysis module will further need to be slightly adapted to be able to treat mixed IF – French input strings.

REFERENCES

[1] Manny Rayner, David Carter, Pierrete Bouillon, Vassilis Digalakis, Mats Wirén “*Spoken Language Translation*” Cambride Press, 2000.

[2] S. Burger, L. Besacier, P. Coletti, F. Metze, C. Morel "*The NESPOLE! VoIP Dialogue Database*", Eurospeech 2001, Aalborg, Danemark, September 2001

[3] L. Besacier, H. Blanchon, Y. Fouquet, J.P. Guilbaud, S. Helme, S. Mazenot, D. Moraru, D. Vaufreydaz "*Speech Translation for French in the NESPOLE! European Project*", Eurospeech 2001, Aalborg, Danemark, September 2001

[4] Levin L. & al. *An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues*. Proc. ICSLP'98, 30th November - 4th December 1998, Sydney, Australia, vol.4/7, pp.1155- 1158.

[5] Blanchon, H. (2002). *A Pattern-Based Analyzer for French in the Context of Spoken Language Translation: First Prototype and Evaluation*. Proc. COLING. Taipei, Taiwan. 24 August - 1 September, 2002.

[6] Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., and Ward, W. "*Recent Advances in JANUS: A Speech Translation System*". Eurospeech, 1993, volume 2, pages 1295-1298.

[7] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai anh Robert L. Mercer. 1992. “*Class-based n-gram Models of Natural Language. Computational Linguistics*” 18(4): 467-479.

[8] Reinherd Kneser and Hermann Ney. 1993. *Improved Clustering Techniques for class-based Statistical Language Modelling*. “In proceeding of the 3rd European Conference on Speech Communication and Technology”, 973-976

[9] Takuya Matsuzaki, Yusuke Miyao, Jun’ichi Tsujii. “*An efficient clustering algorithm for class-based language models*”.