IMPROVING SPOKEN LANGUAGE UNDERSTANDING WITH INFORMATION RETRIEVAL AND ACTIVE LEARNING METHODS

Isabelle Jars, Franck Panaget

France Telecom Research and Development – TECH/EASY
2, avenue Pierre Marzin, 22300 Lannion, France
{isabelle.jars, franck.panaget}@orange-ftgroup.com

ABSTRACT

In the context of deployed spoken dialogue telecom services, we introduce a preprocessor called Fiction into the Spoken Language Understanding (SLU) component. It acts as an intermediate between the speech recognition and interpretation processes in order to increase the rate of utterances that are correctly rejected (CRR for Correctly Rejected Rate) without decreasing the rate of appropriately interpreted utterances. This component is based on statistical approaches of natural language treatment and contextual information. We also use active learning methods to determine the best training corpus size. On a deployed test corpus, the CRR increases from 60% to 86% and active learning method's results show that better performance can be achieved using fewer training data.

Index Terms— Natural languages, Speech communication, Learning systems, Man-machine systems.

1. INTRODUCTION

User-friendliness and usability of spoken natural language systems is still a main challenge. Applications based on such systems are currently deployed but their semantic domain and dialogue strategies are still limited. For instance, the 3000 service, our first deployed vocal service exploiting natural language technologies, was launched for general public use in October 2005. 3000 is an automatic voice agency that enables customers to obtain information and purchase almost 30 different services (e.g. check their consumption, pay their bills and access the management of their services such as call transferring or voice messaging). The 1014 service is an experimental application which has been tested with real users. It handles residential customer requests and routes calls appropriately to one of the following: service subscription and management, building of customer loyalty, payment problems, Internet dysfunction or line connection problems, and automatic services like credit card payments.

We collected two real user corpora: 6,398 utterances in June 2006 with 1014 and 3,374 utterances in May 2007 with 3000. Each utterance was transcribed and annotated with:

- a predicate that represents its semantic interpretation when it is an In-Domain (ID) utterance with interpretation (e.g. *Payment(bill,creditCard)* means that the user wants to pay her/his bill by credit card), or
- REJECTION when it is an Out-Of-Domain (OOD) utterance. In our case, non-speech detection, comments about the system, third-party conversation and in-domain utterances without interpretation.

We observed that the number of OOD utterances is very high compared to ID utterances: OOD represents 26.3% for 3000 and 14.6% for 1014. We also evaluated the performance of our automatic speech recognition (ASR) and baseline spoken language understanding (SLU) components (SLU input = 1-Best ASR hypothesis) and obtained better results on ID utterances than on OOD utterances in terms of interpretation errors: F1 measures are respectively 0.85 and 0.71 for 3000 and 0.84 and 0.73 for 1014.

On the basis of these observations and with the aim of increasing user satisfaction, we have decided to improve the effectiveness of these services. Our goal is first to increase the rejection rate of OOD utterances without decreasing the interpretation rate of ID utterances.

This paper is organized as follows. In section 2, we describe our preprocessor called Fiction (Filter for Improvement of Concept recogniTION) that is introduced into the SLU component. In section 3, we briefly present an active learning method in order to determine the best training corpus which optimizes our model's results. Section 4 shows evaluation results.

2. APPROACH: CHARACTERISTICS OF FICTION

We aim for our spoken dialog system to interpret the meanings of user's utterances and give appropriate responses to requests (Figure 1). This system is composed of an ASR module, an SLU unit and a dialogue manager. The ASR translates user's acoustic signal into text and

passes that to the SLU unit. The SLU unit forms a semantic representation (in our applications, a predicate) of the utterance and gives it to the dialogue manager which reacts to the semantic input with the appropriate actions to take in response to the user's query.

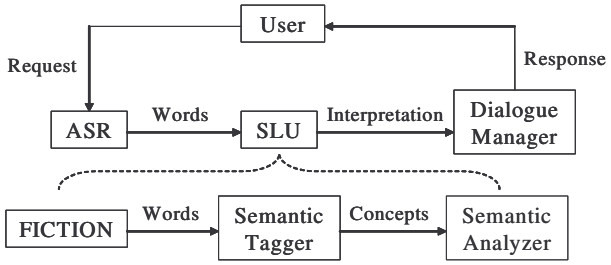

Fig. 1. Our spoken dialogue system.

The SLU unit functionality can be decomposed into two processing steps. The semantic tagger takes the output given by the ASR unit and maps each utterance onto a set of semantic concepts. Then, the semantic analyzer builds the complete semantic interpretation of utterances by applying interpretation rules on concepts.

We suggest adding a component (called Fiction) before the semantic tagger in order to clean and categorize ASR's hypothesis. When Fiction characterizes the ASR's result as ID, the user's utterance is eventually "cleaned" by Fiction then semantically tagged and analyzed in order to produce a semantic representation of the user's utterance (in our applications, a predicate). When ASR's output is categorized as OOD, SLU's result is REJECTION. Unlike different approaches that improve speech recognition (e.g., [1, 2]), Fiction respects word order in sentences. Moreover, it is not only a classifier (such as in [3]) but it can remove one or several words from the ASR's hypotheses. Fiction combines two statistical measures and contextual information. It allows us to determine degrees of association (such as those defined in [4]) between "content word" pairs inside an utterance and between the last utterance and the previous ones. It focuses on content words just as in information retrieval approaches for two main reasons: first, "stop words" have a very low discrimination value in a sentence and co-occur with most words [5] and second, in our applications, concept production and utterance comprehension are mainly based on content words.

Let C be a transcribed training corpus. A user's utterance, U, is a list of content words delimited by <B> and <E>, which represent the beginning and end of U respectively. Let lex(C) be the lexicon of content words that appear in C, plus <B> and <E>. The principle of Fiction is to evaluate the plausibility Plaus(P) where P is a pair of two consecutive words (w_i, w_{i+1}) from U. Note also that the number of times that a word pair P, occurs in C is given by occur(P,C).

Case 1: $w_i, w_{i+1} \in \text{lex}(C) \wedge \text{occur}((w_i, w_{i+1}), C) > 0$

After comparing the results of different likelihood ratios (Dunning, mutual information, T-score, Z-score and Dice coefficient) in our application domain, we have concluded that Dunning's ratio [6] is the best measure to represent co-occurrence of rare events with low frequency of appearance but that are usually significant in a user's sentence. By modeling a unit occurrence as a binomial distribution, Dunning deduces an index which evaluates the plausibility of the independence hypothesis of occurrences of two units. For two words w_j and w_k, Dunning creates a contingency table which takes into account bigrams in which:

- The first word is w_j and the second word is w_k (a);
- The first word is w_j and the second word is not w_k (b);
- The first word is not w_j and the second word is w_k (c);
- The first word is not w_j and the second word is not w_k (d).

We define the plausibility between two content words as equivalent to **Dunning's ratio**. Dunning ratio between words w_j and w_k is calculated as:

$$\text{DunR}(w_j, w_k) = a\log(a) + b\log(b) + c\log(c) + d\log(d) - (a+b)\log(a+b) - (a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) + (a+b+c+d)\log(a+b+c+d)$$

Let us define the minimal value of the Dunning's ratio for all word pairs that appear in a corpus C as follow:

$$DunRMin(C) = \underset{w_j, w_k \in lex(C) \wedge occur((w_j, w_k), C) > 0}{\arg\min} (DunR(w_j, w_k))$$

Case 2: $w_i, w_{i+1} \in \text{lex}(C) \wedge \text{occur}((w_i, w_{i+1}), C) = 0$

Fiction estimates a plausibility measure based on Dagan's similarity formula in which the mutual information is replaced by the Dunning ratio (when defined). We use this measure because we believe that no such word combinations occur in any given corpus even if the corpus is voluminous [7]. Similarity is evaluated in the following way:

$$Sim(w_j, w_k) = \frac{1}{2|lex(C)|} \sum_{x \in lex(C)} \frac{\min(I(x, w_j), I(x, w_k))}{\max(I(x, w_j), I(x, w_k))} + \frac{\min(I(w_j, x), I(w_k, x))}{\max(I(w_j, x), I(w_k, x))}$$

with

$$I(x, y) = \begin{cases} DunR(x, y) & if \quad occur((x, y), C) > 0 \\ 0 & otherwise \end{cases}$$

Fiction decides which word will be removed by comparing this similarity with the minimum value of Dunning's ratio. Two cases are accounted for:

Case 2.1: $-\log(\text{Sim}(w_i, w_{i+1})) \geq \alpha\,\text{DunRMin}(C)$

The words pair, P, is considered plausible and we note that $\text{Plaus}(w_i, w_{i+1}) = -\log(\text{Sim}(w_i, w_{i+1}))$.

Case 2.2: $-\log(\text{Sim}(w_i, w_{i+1})) < \alpha\,\text{DunRMin}(C)$

The word w_i or w_{i+1} will be removed by Fiction according to different values of i:

- If w_i = <B> (i.e. i=0), Fiction removes w_{i+1}.
- If w_{i+1} = <E> (i.e. i=n), Fiction removes w_i.

- If $w_i \neq <B> \wedge w_{i+1} \neq <E>$ (i.e. $i \in [1..n-1]$) , Fiction removes the word with the lowest plausibility with w_{i-1} (Plaus(w_{i-1}, w_i) , Plaus(w_{i-1}, w_{i+1})).

Case 3: $w_i \notin \text{lex}(C) \vee w_{i+1} \notin \text{lex}(C)$

Fiction is not able to take a decision so the words pair is considered plausible and we note: Plaus(w_i, w_{i+1}) = ε

Fiction also uses words from previous utterances, i.e., **dialogue history,** in its algorithm. This information is particularly important for short utterances in which not all content words aren't co-occurrent and for utterances whose nearly all content words were removed (see case 2.2 above). It allows us either to choose content words (of an utterance) that have a sufficient plausibility with other words in the dialogue history (e.g. same subject during the discourse) or to reject the utterance (REJECTION). Note that this information can't be exploited for all applications because some of them have few interactions with the user to build a robust context.

To demonstrate our component's workings, we briefly show three examples of Fiction's results compared to ASR output and real user's utterance which are extracted from the 3000 service corpus (and translated from French). The first and second examples show that our component can delete one word or part of user's utterance which aren't plausible. The third example presents Fiction's result when all of the utterance is rejected.

1. *User*: *I want to pay my phone hum bill*
   ASR: I want to pay my phone mail bill
   FICTION: I want to pay my phone bill
2. *User*: [Private conversation: *Turn the music down*] *I want to pay my phone bill*
   ASR: I want servers I want to pay my phone bill
   FICTION: I want to pay my phone bill
3. *User*: [Private conversation: *Put your shoes on Charlène*]
   ASR: but if I phone
   FICTION: REJECTION

3. ACTIVE LEARNING METHOD FOR OPTIMIZING TRAINING CORPUS' SIZE

As we have described, we use statistical approaches to clean and categorize ASR's outputs. In order to constitute the best training corpus, we seek to reduce the number of training examples by selecting those which will have the largest improvement on Fiction's performance. Inspired by certainty-based active learning methods [8, 9], the selection of training examples is also based on Dunning's ratio.

For that, we utilize the transcribed training corpus used to build ASR's linguistic model. The corpus is divided into a small amount of transcribed data C_0, a development corpus D and a large amount of remaining transcribed data C. We build Fiction's model M(C_0) and calculate the Interpretation Error Rate (IER) on corpus D. As in [1], IER is obtained by summing up the different errors (false alarms, substitutions and false rejections) and dividing by the total amount of non-empty reference interpretation. Then, we build a second model M(C_1) where C_1 is C_0 increased with additional utterances from C. These utterances are selected as follow: 1) we identify content words pairs from C whose ratio is closed to the DunRMin(C_0), and then 2) we extract utterances that contain those content word pairs. The algorithm loops until a model M(C_i) that minimizes the IER (on corpus D) is obtained.

At runtime, Fiction stores all recognized utterances including content words that don't appear in its training corpus. These utterances are good candidates to be transcribed in order to boost, at least, Fiction's model.

4. EXPERIMENTS AND RESULTS

A comparison between ASR+SLU with Fiction (with factor α=1) and ASR+SLU without Fiction (baseline) on both real-users corpora (cf. § 1) is shown in table 1. Results on ID utterances are interpretation rates (but not the performance of categorizing utterances as ID utterances).

	3000			1014		
	Baseline	Fiction	%	Baseline	Fiction	%
IER	24,05%	12,89%	-46,40%	20,26%	18,89%	-6,75%
Recall for ID utt.	0,9033	0,9198	1,83%	0,8468	0,8332	-1,61%
Precision for ID utt.	0,8106	0,9176	13,21%	0,8369	0,8685	3,77%
F1 measure for ID utt.	0,8544	0,9187	7,52%	0,8418	0,8504	1,02%
Recall for OOD utt.	0,5998	0,8643	44,11%	0,7035	0,8676	23,33%
Precision for OOD utt.	0,8799	0,8702	-1,11%	0,7574	0,6974	-7,91%
F1 measure for OOD utt.	0,7133	0,8673	21,58%	0,7294	0,7733	6,01%

Table 1. Results for applications 3000 and 1014.

We detect an increase of utterances that are correctly rejected in both applications (+44% for 3000 and +23% for 1014) without too much decrease in the interpretation rate of ID utterances (an increase of 1.8% for 3000). These results are in keeping with our aim because they increase the CRR and also have a considerable benefit in terms of user request comprehension. The F1-measure of OOD utterances is increased by 21% for 3000 and 6% for 1014, with a worsening of 1% for 3000 and 8% for 1014 for the precision. Precision diminution is essentially due to a false rejection rise in our model, which should be investigated. Nevertheless, we prefer to increase an utterance rejection instead of losing trust in ASR's results (due to an impossible similarity measure). In fact, according to an internal ergonomic study produced in 1999 which shows that only 50% of users immediately repair a system's misunderstandings, we wish to favor user's repetitions over user's repairs. Note that the results given for 1014 are not as significant as 3000 because we used a small training corpus which was not optimized. These first results are experimental for this application because we are only beginning to work with these data.

We have evaluated our active learning method on the 3000 application. The training corpus is made up of 53,569 utterances (C_0 = 500 utterances, D=3,000, and C=50,069). In order to see the actual improvement, we performed controlled experiments comparing Fiction with active selection, Fiction with random selection, and the baseline system (SLU without Fiction). Figures 2a and 2b show the evolution of F1-measure for ID and ODD utterances respectively. The best performances are obtained with the active selection of a corpus of 5,000 utterances. At this point, Dunning's ratio is optimal for determining the degree of association and similarity between content words in an utterance. Adding new examples in the training corpus decreases the performances; a similar phenomenon is observed in [8]. We have succeeded in almost reducing the IER percent by half (13% with a 5,000 utterance corpus against 24% for the baseline). Note that because these are controlled experiments, the performance using all available training data is the same both for random selection and the active learning method.

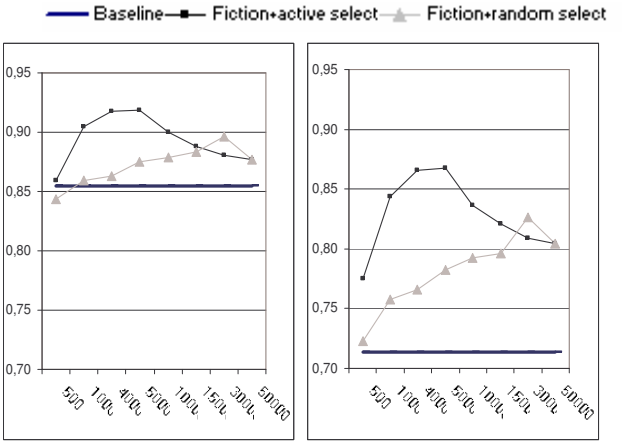

Fig. 2a. Evolution of F1-measure on ID utterances.

Fig. 2b. Evolution of F1-measure on OOD utterances.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a new method to increase the rate of utterances that are correctly rejected by combining statistical approaches of natural language treatment and contextual methods. We have also presented an active learning method in order to reduce the number of training examples. Our experiments on two different deployed services show that results are significant with respect to our goals. We have determined that the rate of utterances that are correctly rejected increases for interpretation. Therefore, these first results bring substantial benefits in terms of user request comprehension and thus reduce the time invested by the user with vocal services. We then proved that our active learning method can be viewed as an optimized algorithm which selects the best training corpus. We also demonstrate that using active learning for selecting training data is better than using random selection because the CRR maximized more quickly and with less words. This conclusion is particularly appealing because it demonstrates that it isn't necessary to have a large corpus for obtaining the best results.

For future work, we hope to increase the effectiveness and importance of our active learning method by selecting training corpus on untranscribed data and combining it with active learning methods used for building ASR's training corpus. That would reduce the time and costs of corpus acquisition and annotation.

6. ACKNOWLEDGEMENTS

This work is supported by the 6th Framework Research Programme of the European Union (EU), Project LUNA, IST contract n° 33549. The authors would like to thank the EU for the financial support. For more information about the LUNA project, please visit http://www.ist-luna.eu.

7. REFERENCES

[1] E. Filisko and S. Seneff, "Error Detection and Recovery in Spoken Dialogue Systems," *Proceedings of Workshop for Spoken Language Understanding for Conversational Systems*, 2004.

[2] J. Glass, "Challenges for spoken dialogue systems," in *Proceedings of IEEE ASRU Workshop*, 1999.

[3] B. Minescu, G. Damnati, F. Béchet, and R. De Mori, "Conditional use of word lattices, confusion networks and 1-best string hypotheses in a sequential interpretation strategy," in *Proceeding of Interspeech*, 2007.

[4] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[5] R. Tsz-Wai Lo, B. He and L. Ounis, "Automatically Building a Stop Word List for an Information Retrieval System," *Journal of digital information management*, vol. 3(4), pp. 3-9, 2005.

[6] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19(1), pp.61-74, 1993.

[7] I. Dagan, L. Lee and F.C.N Pereira, "Similarity-based models of word co-occurrence probabilities," *Machine learning*, vol. 34(1-3), pp.43-69, 1999.

[8] G. Tür, R.E., Schapire and D. Hakkani-Tür, "Active learning for spoken language understanding," in *Proceeding of the ICASSP*, 2003.

[9] D.D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proceedings of the ICML*, 1994.