Using Paraphrases and Lexical Semantics to Improve the Accuracy and the Robustness of Supervised Models in Situated Dialogue Systems
Claire Gardent, Lina Maria Rojas Barahona

To cite this version:
Claire Gardent, Lina Maria Rojas Barahona. Using Paraphrases and Lexical Semantics to Improve the Accuracy and the Robustness of Supervised Models in Situated Dialogue Systems. Conference on Empirical Methods in Natural Language Processing, SIGDAT, the Association for Computational Linguistics special interest group on linguistic data and corpus-based approaches to NLP, Oct 2013, Seattle, United States. pp.808-813. hal-00905405

HAL Id: hal-00905405
https://inria.hal.science/hal-00905405v1
Submitted on 18 Nov 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Using Paraphrases and Lexical Semantics to Improve the Accuracy and the Robustness of Supervised Models in Situated Dialogue Systems

Claire Gardent  
CNRS/LORIA, Nancy  
claire.gardent@loria.fr

Lina M. Rojas Barahona  
Université de Lorraine/LORIA, Nancy  
lina.rojas@loria.fr

Abstract

This paper explores to what extent lemmatisation, lexical resources, distributional semantics and paraphrases can increase the accuracy of supervised models for dialogue management. The results suggest that each of these factors can help improve performance but that the impact will vary depending on their combination and on the evaluation mode.

1 Introduction

One strand of work in dialog research targets the rapid prototyping of virtual humans capable of conducting a conversation with humans in the context of a virtual world. In particular, question answering (QA) characters can respond to a restricted set of topics after training on a set of dialogs whose utterances are annotated with dialogue acts (Leuski and Traum, 2008).

As argued in (Sagae et al., 2009), the size of the training corpus is a major factor in allowing QA characters that are both robust and accurate. In addition, the training corpus should arguably be of good quality in that (i) it should contain the various ways of expressing the same content (paraphrases) and (ii) the data should not be skewed. In sum, the ideal training data should be large (more data is better data); balanced (similar amount of data for each class targeted by the classifier) and varied (it should encompass the largest possible number of paraphrases and synonyms for the utterances of each class).

In this paper, we explore different ways of improving and complementing the training data of a supervised QA character. We expand the size and the quality (less skewed data) of the training corpus using paraphrase generation techniques. We compare the performance obtained on lemmatised vs. non lemmatised data. And we investigate how various resources (synonym dictionaries, WordNet, distributional neighbours) can be used to handle unseen words at run time.

2 Related work

Previous work on improving robustness of supervised dialog systems includes detecting and handling out of domain utterances for generating feedback (Lane et al., 2004); using domain-restricted lexical semantics (Hardy et al., 2004); and work on manual data expansion (DeVault et al., 2011). Our work follows up on this research but provides a systematic investigation of how data expansion, lemmatisation and synonym handling impacts the performance of a supervised QA engine.

3 Experimental Setup

We run our experiments on a dialog engine developed for a serious game called Mission Plastech Technologie. In this game, the player must interact with different virtual humans through a sequence of 12 subdialogs, each of them occurring in a different part of the virtual world.

Training Data. The training corpus consists of around 1250 Human-Human dialogues which were manually annotated with dialog moves. As the following dialog excerpt illustrates, the dialogs are conducted in French and each dialog turn is manually annotated using a set of 28 dialog acts. For
a more detailed presentation of the training corpus and of the annotation scheme, the reader is referred to (Rojas-Barahona et al., 2012a)

dialog : 01_dialogDirecteur-Tue Jun 14 11:04:23 2011
>M.Jasper : Bonjour, je suis M.Jasper le directeur. \textit{greet} (Hello, I am the director, Mr. Jasper.)
>M.Jasper : Qu’est-ce que je peux faire pour vous ? \textit{ask(task(X))} (What can I do for you?)
>Lucas : je dois sauver mon oncle \textit{first_step} (I must rescue my uncle)
>M.Jasper : Pour faire votre manette, il vous faut des plans. Allez voir dans le bureau d’études, ils devraient y être. \textit{inform(do(first_step))} (To build the joystick you will need the plans. You will find them in the Designing Office.)
>M.Jasper : Bonne Chance ! \textit{quit} (Good Luck!)

\textit{Dialog Systems} For our experiments, we use a hybrid dialog system similar to that described in (Rojas-Barahona et al., 2012b; Rojas Barahona and Gardent, 2012). This system combines a classifier for interpreting the players utterances with an information state dialog manager which selects an appropriate system response based on the dialog move assigned by the classifier to the user turn. The classifier is a logistic regression classifier\(^1\) which was trained for each subdialog in the game. The features used for training are the set of content words which are associated with a given dialog move and which remain after TF*IDF\(^2\) filtering. Note that in this experiment, we do not use contextual features such as the dialog acts labeling the previous turns. There are two reasons for this. First, we want to focus on the impact of synonym handling, paraphrasing and lemmatisation on dialog management. Removing contextual features allows us to focus on how content features (content words) can be improved by these mechanisms. Second, when evaluating on the H-C corpus (see below), contextual features are often incorrect (because the system might incorrectly interpret and thus label a user turn). Excluding contextual features from training allows for a fair comparison between the H-H and the H-C evaluation.

\textit{Test Data and Evaluation Metrics} We use accuracy (the number of correct classifications divided by the number of instances in the testset) to measure performance and we carry out two types of evaluation. On the one hand, we use 10-fold cross-validation on the EmoSpeech corpus (H-H data). On the other hand, we report accuracy on a corpus of 550 Human-Computer (H-C) dialogues obtained by having 22 subjects play the game against the QA character trained on the H-H corpus. As we shall see below, performance decreases in this second evaluation suggesting that subjects produce different turns when playing with a computer than with a human thereby inducing a weak out-of-domain effect and negatively impacting classification. Evaluation on the H-H corpus therefore gives a measure of how well the techniques explored help improving the dialog engine when used in a real life setting.

Correspondingly, we use two different tests for measuring statistical significance. In the H-H evaluation, significance is computed using the Wilcoxon signed rank test because data are dependent and are not assumed to be normally distributed. When building the testset we took care of not including paraphrases of utterances in the training partition (for each paraphrase generated automatically we keep track of the original utterance), however utterances in both datasets might be generated by the same subject, since a subject completed 12 distinct dialogues during the game. Conversely, in the H-C evaluation, training (H-H data) and test (H-C data) sets were collected under different conditions with different subjects therefore significance was computed using the McNemar sign-test (Dietterich, 1998).

4 Paraphrases, Synonyms and Lemmatisation

We explore three main ways of modifying the content features used for classification: lemmatising the training and the test data; augmenting the training data with automatically acquired paraphrases; and substituting unknown words with synonyms at run time.

\textbf{Lemmatisation} We use the French version of Treetagger\(^3\) to lemmatise both the training and the test data. Lemmas without any filtering were used

---

1. We used MALLET (McCallum, 2002) for the LR classifier with L1 Regularisation.
2. TF*IDF = Term Frequency*Inverse Document Frequency
3. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
to train classifiers. We then compare performance with and without lemmatisation. As we shall see, the lemma and the POS tag provided by TreeTagger are also used to lookup synonym dictionaries and EuroWordNet when using synonym handling at run time.

**Paraphrases:** (DeVault et al., 2011) showed that enriching the training corpus with manually added paraphrases increases accuracy. Here we exploit automatically acquired paraphrases and use these not only to increase the size of the training corpus but also to better balance it\(^4\). We proceed as follows.

First, we generated paraphrases using a pivot machine translation approach where each user utterance in the training corpus (around 3610 utterances) was translated into some target language and back into French. Using six different languages (English, Spanish, Italian, German, Chinese and Arabian), we generated around 38000 paraphrases. We used Google Translate API for translating.

---

Table 1: Skewed and Balanced Data on a sample sub-dialog. The category with lowest number of paraphrases is **greet**, with 62 paraphrases, hence \( l_p = 62 \). All categories were increased by 62 except **find_plans** and **yes** that were increased by half: 31.

| Category        | Train Instances | Balanced Instances |
|-----------------|-----------------|--------------------|
| greet           | 24              | 86                 |
| help            | 20              | 82                 |
| yes             | 92              | 123                |
| no              | 55              | 117                |
| ack             | 73              | 135                |
| other           | 27              | 89                 |
| quit            | 38              | 100                |
| find_plans      | 115             | 146                |
| job             | 26              | 88                 |
| staff           | 15              | 77                 |
| studies         | 20              | 82                 |
| security_policies | 24         | 86                 |
| \( \mu \)       | 44.08           | 100.92             |
| \( \sigma \)    | ±32.68          | ±23.32             |

---

\( l_p \) is supposed to ask information about the joystick in the Emospeech game. In this dialogue the player is supposed to ask information about the joystick plans (**find_plans**, which is the mandatory goal). The categories cover mandatory and optional goals and general dialogue acts, such as greetings, asking for help, confirm and disconfirm, acknowledgment and out of topic questions (i.e. other).

**Substituting Synonyms for Unknown Words** A word is unknown, if it is a well-formed French

---

\( l_p \) is supposed to ask information about the joystick in the Emospeech game. In this dialogue the player is supposed to ask information about the joystick plans (**find_plans**, which is the mandatory goal). The categories cover mandatory and optional goals and general dialogue acts, such as greetings, asking for help, confirm and disconfirm, acknowledgment and out of topic questions (i.e. other).

**Substituting Synonyms for Unknown Words** A word is unknown, if it is a well-formed French
word and if it does not appear in the training corpus. Conversely, a word is known if it is not unknown.

When an unknown word $w$ is detected in a player utterance at runtime, we search for a word $w'$ which occurs in the training data and is either a synonym of $w$ or a distributional neighbour. After disambiguation, we substitute the unknown word for the synonym.

To identify synonyms, we make use of two lexical resources namely, the French version of EuroWordNet (EWN) (Vossen, 1998), which includes 92833 synonyms, hyperonyms and hyponyms pairs, and a synonym lexicon for French (DIC) which contains 38505 lemmas and 254149 synonym pairs. While words are categorised into Noun, Verbs and Adjectives in EWN, DIC contains no POS tag information.

To identify distributional neighbours, we constructed semantic word spaces for each subdialog in the EmoSpeech corpus using random indexing (RI) on the training corpus expanded with paraphrases. Using the cosine measure as similarity metrics, we then retrieve for any unknown word $w$, the word $w'$ which is most similar to $w$ and which appears in the training corpus.

For lexical disambiguation, two methods are compared. We use the POS tag provided by TreeTagger. In this case, disambiguation is syntactic only. Or we pick the synonym with highest probability based on a trigram language model trained on the H-H corpus.

5 Results and Discussion

Table 2 summarises the results obtained in four main configurations: (i) with and without paraphrases; (ii) with and without synonym handling; (iii) with and without lemmatisation; and (iv) when combining lemmatisation with synonym handling. We also compare the results obtained when evaluating using 10-fold cross validation on the training data (H-H dialogs) vs. evaluating the performance of the system on H-C interactions.

**Overall Impact** The largest performance gain is obtained by a combination of the three techniques explored in this paper namely, data expansion, synonym handling and lemmatisation (+8.9 points for the cross-validation experiment and +2.3 for the H-C evaluation).

**Impact of Lexical Substitution at Run Time** Because of space restrictions, we do not report here the results obtained using lexical resources without lemmatisation. However, we found that lexical resources are only useful when combined with lemmatisation. This is unsurprising since synonym dictionaries and EuroWordNet only contain lemmas. Indeed when distributional neighbours are used, lemmatisation has little impact (e.g., 65.11% using distributional neighbours without lemmatisation on the H-H corpus without paraphrases vs. 66.41% when using lemmatisation).

Another important issue when searching for a word synonym concerns lexical disambiguation: the synonym used to replace an unknown word should capture the meaning of that word in its given context. We tried using a language model trained on the training corpus to choose between synonym candidates (i.e., selecting the synonym yielding the highest sentence probability when substituting that synonym for the unknown word) but did not obtain a significant improvement. In contrast, it is noticeable that synonym handling has a higher impact when using EuroWordNet as a lexical resource. Since EuroWordNet contain categorial information while the synonym dictionaries we used do not, this suggests that the categorial disambiguation provided by TreeTagger helps identifying an appropriate synonym in EuroWordNet.

Finally, it is clear that the lexical resources used for this experiment are limited in coverage and quality. We observed in particular that some words which are very frequent in the training data (and thus which could be used to replace unknown words) do not occur in the synonym dictionaries. For instance when using paraphrases and dictionaries (fourth row and
Acknowledgments

This work was partially supported by the EU funded Eurostar EmoSpeech project. We thank Google for giving us access to the University Research Program of Google Translate.
References

David DeVault, Anton Leuski, and Kenji Sagae. 2011. Toward learning and evaluation of dialogue policies with text examples. In 12th SIGdial Workshop on Discourse and Dialogue, Portland, OR, June.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10:1895–1923.

Hilda Hardy, Tomek Strzalkowski, Min Wu, Cristian Ursu, Nick Webb, Alan W. Biermann, R. Bryce Inouye, and Ashley McKenzie. 2004. Data-driven strategies for an automated dialogue system. In ACL, pages 71–78.

Ian Richard Lane, Tatsuya Kawahara, and Shinichi Ueno. 2004. Example-based training of dialogue planning incorporating user and situation models. In INTERSPEECH.

Anton Leuski and David Traum. 2008. A statistical approach for text processing in virtual humans. In Proceedings of the 26th Army Science Conference.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Lina Maria Rojas Barahona and Claire Gardent. 2012. What should I do now? Supporting conversations in a serious game. In SeineDial 2012 - 16th Workshop on the Semantics and Pragmatics of Dialogue, Paris, France, Jonathan Ginzburg (chair), Anne Abeillé, Margot Colinet, Gregoire Winterstein.

Lina M. Rojas-Barahona, Alejandra Lorenzo, and Claire Gardent. 2012a. Building and exploiting a corpus of dialog interactions between french speaking virtual and human agents. In Proceedings of the 8th International Conference on Language Resources and Evaluation.

Lina M. Rojas Barahona, Alejandra Lorenzo, and Claire Gardent. 2012b. An end-to-end evaluation of two situated dialog systems. In Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 10–19, Seoul, South Korea, July. Association for Computational Linguistics.

K. Sagae, G. Christian, D. DeVault, , and D.R. Traum. 2009. Towards natural language understanding of partial speech recognition results in dialogue systems. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Companion Volume: Short Papers, pages 53–56.

Benoît Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In 7th international conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta.

Piek Vossen, editor. 1998. EuroWordNet: a multilingual database with lexical semantic networks. Kluwer Academic Publishers, Norwell, MA, USA.