AUTOMATIC DIALOG ACT CORPUS CREATION FROM WEB PAGES

Pavel Král
Department of Computer Science and Engineering, University of West Bohemia, Plzeň, Czech Republic

Christophe Cerisara
LORIA UMR 7503, BP 239 - 54506 Vandoeuvre, France

Keywords: Automatic labeling, Corpus, Dialog act, Internet.

Abstract: This work presents two complementary tools dedicated to the task of textual corpus creation for linguistic researches. The chosen application domain is automatic dialog acts recognition, but the proposed tools might also be applied to any other research area that is concerned with dialogs processing. The first software captures relevant dialogs from freely available resources on the World Wide Web. Filtering and parsing of these web pages is realized thanks to a set of hand-crafted rules. A second set of rules is then applied to achieve automatic segmentation and dialog act tagging. The second software is finally used as a post-processing step to manually check and correct tagging errors when needed. In this paper, both softwares are presented, and the performances of automatic tagging are evaluated on a dialog corpus extracted from an online Czech journal. We show that reasonably good dialog act labeling accuracy may be achieved, hence greatly reducing the cost of building such corpora.

1 INTRODUCTION

Modeling and automatically identifying the structure of spontaneous dialogs is very important to better interpret and understand them. The precise modeling of spontaneous dialogs is still an open issue, but several specific characteristics of dialogs have already been clearly identified. Dialog Acts (DAs) are one of these characteristics.

Dialog acts are defined by Austin in (Austin, 1962) as the meaning of an utterance at the level of illocutionary force. In other words, the dialog act is the function of a utterance (or its part) in the dialog. For example, the function of a question is to request some information, while an answer shall provide this information.

Dialog acts are also used in Spoken Language Understanding. In this area, dialog acts are defined much more precisely, but they are also often application-dependent. Hence, Jeong *et al.* define in (Jeong and Lee, 2006) a dialog act as a domain-dependent intent, such as “Show Flight” or “Search Program”, respectively in the flight reservation and electronic program guide domains.

One of the main issue in the automatic dialog acts recognition field concerns the lack of training data and the design of fast and cheap methods to create new corpora.

The main goal of this work is three-fold:

1. Design semi-automatic procedures to build such corpora at a low cost.
2. Develop dedicated tools that implement these procedures.
3. Assess their performances on the concrete task of building a textual corpus in the Czech language annotated with dialog acts.

Two main steps are required to actually build a corpus:

1. Data acquisition.
2. Data labeling.

Both steps can be realized manually in order to guarantee the best possible quality of the corpus, but this is then extremely costly and time consuming. Most often the first step of data acquisition is realized based on existing resources. Nowadays, a great amount of information and textual content is available

Král P. and Cerisara C. (2010).
AUTOMATIC DIALOG ACT CORPUS CREATION FROM WEB PAGES.
In *Proceedings of the 12th International Conference on Enterprise Information Systems - Human-Computer Interaction*, pages 198-203
DOI: 10.5220/0003019501980203

on the Web. Among these resources, many dialogs in many different languages are available. Therefore, a lot of efforts has been put in order to exploit the World Wide Web as a primary content source for corpus building. The proposed work follows the same strategy and also focuses on extracting textual online content for the target application. Yet, online texts are difficult to exploit because of the wide variability of encoding and representation formats, which makes the extraction process challenging. Note, that in this work, only texts are extracted, and the audio recordings are discarded for the time being. We have indeed shown in previous works (Král et al., 2006), (Král et al., 2007), (Král et al., 2008) that textual transcriptions are the most informative features for dialog act recognition, which justifies this choice as a first approximation.

The rest of the paper is organized as follows. The next section presents a short review of corpus creation approaches. Section 3 describes the processing of Web pages along with the proposed approach for automatic segmentation and dialog act labeling. Section 4 describes the *jDALabeler* tool for manual annotation and correction, while Section 5 evaluates the whole process on building a Czech corpus. In the last section, we discuss these results and propose some future research directions.

2 SHORT REVIEW OF CORPORA CREATION

Because of the virtually unlimited amount of textual content freely available on the World Wide Web, many research works have tried to extract and exploit useful information from the Web in the last years, in order to create corpora for a variety of applications and domains. We only review next a few of these works that are closely related to our application or that illustrate some specific and important benefits of exploiting these kinds of resources.

Maeda *et al* present in (Maeda et al., 2008) several tools/systems for various corpus creation projects. The following main issues are addressed:

- Data scouting: to browse Web pages and save them in a database.
- Data selection: to chose the adequate data source.
- Data annotation: to associate the data with its corresponding labels.
- Speech transcription.

Most often, these processes are realized manually, which guarantees a good quality but greatly increases the overall cost. A major advantage of extracting information from Web pages is that textual resources are often completed with contextual information, such as keywords, document summary, related videos, and so on. However, most of this information is stored in different and non-standard formats, requiring advanced methods, such as automatic classifiers, to filter and "interpret" them. For instance, (Zhang et al., 2006) exploit k-nearest-neighbor classifiers to grab bilingual parallel corpora from the Web.

The work described in (Sarmento et al., 2009) is more closely related to our work; it also exploits fine-tuned hand-crafted rules to classify sentences. However, it mainly differs on the chosen application - political opinions mining, on the chosen language - Portuguese, on the textual sources - online newspapers and on the manual rules. Many research efforts in the domain of corpus creation are actually dedicated to gathering and compiling available resources in a given language, for which large enough corpora may not exist yet (Pavel et al., 2006). The Web is very important in these aspects, and several such works thus focus on comparing and exploring the most efficient ways to crawl the web and retrieve relevant information (Botha and Barnard, 2005).

3 WEB PAGES PROCESSING AND AUTOMATIC CORPUS LABELING

The *MOIS* (Monitoring Internet Resources) software is a specific Web crawler designed to process Web pages with dialogs in order to build new corpora. The algorithm for processing a website is the following:

1. Start from a given URL.
2. Detect, clean and save all dialogs in the Web page corresponding to this URL.
3. Parse the dialogs: segment into sentences and annotate all sentences with dialog act tags.
4. Store all hyperlinks in this Web page into a list.
5. Chose (and remove from the list) one of the saved URL and iterate from step 1 until the depth of the website exceeds *n*.

3.1 Dialog Detection

During step 2, dialog detection and extraction is achieved based on hand-made rules. These rules exploit several features, such as:

- Information about the speakers (e.g. speaker identity).

- Regular alternation of different font styles of paragraphs (e.g. bold, normal).
- Regular paragraph separation by vertical bars.
- Alternation of modalities (e.g. succession of sentences with a final "." punctuation mark interleaved with sentences with a final question mark).

Dialogs are then cleaned by removing all *html* and *xml* tags from the Web pages: only "raw" text with punctuation is passed to the next processing steps.

Without any prior information about their internal structure, it is very difficult to completely clean all Web pages because of the variety of websites' structures. Hence, a subsequent manual checking and correction may be sometimes needed, and this functionality should be available in the annotation software.

In order to minimize these corrections, the first version of *MOIS* is designed to specifically support an online Czech journal called "Super" (see `http://www.super.cz/svet-celebrit/rozhovory`), which contains many dialogs. Hence, in its most generic version, *MOIS* processes websites without any prior information about their structure, but a specific HTML parsing has been implemented for the case of the website of the journal "Super".

3.2 Sentence Segmentation and Annotation

Sentence segmentation is realized based on a simple rule, which looks for the sequences of characters $end\_sentence\_mark + space + capital\_letter$ where *end_sentence_mark* can be ".", "!", ":", ";" or "?", and segments the raw text every time such a sequence is detected.

Automatic dialog act labeling also analyses the punctuation marks with a set of lexical and grammar rules. In the first version of *MOIS*, only four elementary DAs are considered: statements (S), exclamations (E), yes/no questions (Q[y/n]) and other questions (Q).

The classification algorithm is then as follows:

- E : sentences with a final "!" mark.
- Q : sentences with a final "?" mark and which start with an interrogative word.
- Q : sentences with a final "?" mark and with an interrogative word just after a comma.
- Q[y/n] : other sentences with a final "?" mark.
- S : all others utterances.

In the second version of *MOIS*, we plan to implement more precise rules, such as detecting the inversion of the subject-verb pair for questions. We will further add more keywords (e.g. acknowledgment detection), as well as several other enhancement of this basic set of rules. The choice of which rules to apply at this stage will be made based on empirical studies of the behavior of the system in this first set of experiments.

3.3 Software Description

MOIS is a Web application based on the J2EE servlets technology. It is composed of two main windows: the first one is used to input the URL to process and to save the resulting dialogs into XML files; the second one is used to show the resulting dialogs.

The main screen of *MOIS* is shown in Figure 1.

Figure 1: Main screen of the *MOIS* tool.

4 MANUAL ANNOTATION CORRECTION

Automatic annotation is not perfect, and manual correction of dialog act labels might be required. In order to perform such a manual annotation, we have developed the *jDALabeler* software, which is a tool dedicated to manual corpus labeling with predefined labels (dialog acts in this case) at the sentence level.

jDALabeler is an OS-independent software developed in the java programming language. This tool contains a graphical user interface, which is controlled by a combination of keyboard shortcuts and mouse controls.

It contains two predefined DA tag-sets that are based on the two very popular DA taxonomies: Meeting Recorder DA (MRDA) tag-set (Dhillon et al., 2004) and VERBMOBIL (Jekat *et al.*, 1995). It is possible to change or complete these predefined tag-sets via a configuration file. The user can also define its own DA taxonomy. The hierarchical DA structure

is represented by a tree that is saved in *xml* format. An example of a part of such a tree is given next:

```
<daGeneralTag name="questions" keyShort Cut="q"
detailExplanation="to request some information">
  <daSpecificTag name="y/nQuestion" keyShortCut="qy"
  detailExplanation="possible answer yes or no"/>
  <daSpecificTag name="whQuestion" keyShortCut="w"
  detailExplanation="contains wh word" />
  ...
</daGeneralTag>
...
```

The corresponding *xml* format is validated with *xsd* schema.

During the labeling process, the user first selects some sentence to label. Then, he may either tag the selected text by choosing the corresponding DA with the mouse or with a keyboard shortcut. By default, plain text files with UTF-8 encoding are supported by the software. Although this tool is primarily designed to support speech recogniser outputs, it may also load textual files stored in *xml* format. The labeled text is finally saved in *xml* format as shown in the following example:

```
<?xml version="1.0" encoding="UTF-8"?>
<document annotationSchemeDA="MRDA">
<speakers>
    <speaker id="1" name="Josef Husa"/>
    ...
</speakers>
<content>
    Ano, tak jsem to myslel a beru to. Jak
    ale chcete vyřešit tohle? ...
</content>
<annotationDA>
<speaker id="1">
<sentence>
<daGeneralTag name="responses">
<daSpecificTag name="positive">
    <token>Ano</token>
    <token>,</token>
    <token>tak</token>
    ...
</daSpecificTag>
</daGeneralTag>
</sentence>
...
</speaker>
...
```

The whole text is automatically segmented into units, called "tokens". The tags "speaker id" and "sentence" respectively inform about the speaker identity and the segmentation into sentences.

The main screen of the *jDALabeler* is shown in Figure 2.

A dedicated window of the graphical interface displays the current line of the transcription file, the eventual error messages and the application help.

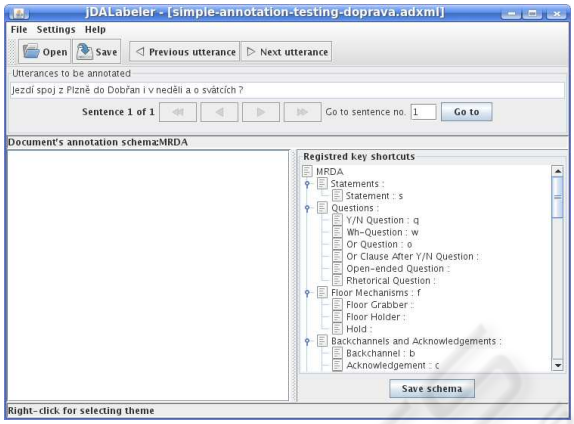

Figure 2: Main screen of the *jDALabeler* tool.

5 DIALOG ACT CORPUS

The tagging functionality of the developed tool *MOIS* has been evaluated on the Czech internet journal "Super", which contains several dialogs between journalists and celebrities. 439 dialogs have been detected and processed. These dialogs have been automatically segmented into sentences and annotated with dialog acts.

As mentioned previously, our dialog act tag-set is composed of four elementary DAs: statements (S), exclamations (E), yes/no questions (Q[y/n]) and other questions (Q).

The composition of the resulting dialog act corpus with illustrative examples of dialog acts is shown in Table 1.

Table 1: Composition of automatically created Czech DA corpus from the journal "Super" with some examples

DA	No.	Example	English translation
S	33567	A firma ho samozřejmě podpoří.	It will be definitely supported by the company.
E	1238	Holky, pojďte ke mě!	Girls, come to me!
Q[y/n]	4135	Počítáte s tím, že školu doděláte?	Do you hope to finish this school?
Q	3156	Kolik prodáváte desek?	How many gramophone records do you sell?
All	42096		

In order to evaluate the quality of the segmentation and labelling of the proposed system, a part of the corpus produced has been manually checked and corrected. One hundred sentences per dialog act class are randomly chosen and their DA labels manually

checked and corrected with *jDALabeler*. Two types of errors are identified: segmentation and labeling errors.

Table 2 shows the ratio of erroneous sentences, respectively for segmentation and classification errors, over the total number of sentences for each dialog act class. The column "Segment." hence reports segmentation errors, while the column "Label." reports DA labeling errors. Labeling errors are further decomposed into two classes, depending on whether these errors may be easy to handle via additional handcrafted lexical/grammatical rules or not. The former errors appear in the "Regular" column, while the latter errors are shown in the "Irregular" column, as they are much more difficult to handle via a simple rule. This clustering is of course subjective, but it also gives some insights about the actual potential of the proposed approach.

Regarding segmentation errors, most of them occur in long sentences with several "," marks and/or sentences that contain textual citations (between double quote symbols).

Table 3 reports the confusion matrix for automatic DA labeling, without segmentation errors. In addition to the four dialog acts described above, a new column "A" is included in this matrix, which corresponds to sentences that should be classified in Another dialog act, such as accepts, rejects, thanks, acknowledgments, etc. These sentences are thus considered as systematic classification errors. We plan to extend our basic set of dialog acts in a future work to handle these sentences.

The following conclusions can be made from error analysis:

- Many errors in class S are sentences that contain speaker identification; i.e. "speaker:".
- Class E is mainly composed of "exclamations", but 5% of them can also be interpreted as orders.
- Class E is the one that contains the most of unsupported DAs, and should thus benefit from extending the basic DA tag set.
- An analysis of confusions in class Q[y/n] shows that many such errors actually correspond to questions that can have two possible answers, other than yes/no; these questions belong for now to class Q, but it may be a good idea to create specific class and rules for these specific questions.
- Conversely, most errors in class Q correspond to confusions with class Q[y/n].

Table 2: Segmentation and classification errors in automatically created DA corpus.

	Error types in [%]			
Class	Segment.	Label.	Regular	Irregular
S	4	9	7	2
E	4	14	0	14
Q[y/n]	2	26	26	0
Q	1	13	12	1

Table 3: Confusion matrix (in %) of automatic DA corpus labeling.

	Automatic labeled class in [%]				
Class	S	E	Q[y/n]	Q	A
S	**91**	0	0	0	9
E	0	**86**	1	1	12
Q[y/n]	0	0	**74**	16	10
Q	0	0	12	**87**	1

6 CONCLUSIONS AND PERSPECTIVES

We have presented in this work a software framework for text corpus creation from the web. An approach for automatic sentence segmentation and dialog act labeling, which exploits a set of manually-defined rules, has also been developed. The performances of this system have been evaluated on a Czech dialog corpus, extracted from an online journal. A part of this corpus has been manually checked to compute the classification accuracy of the chosen lexical and grammatical rules. The resulting confusion matrix has been analyzed, suggesting that the performances obtained with such an automatic approach are good enough to further exploit the extracted data as a bootstrapping corpus for subsequent processing with model-based approaches, such as the ones described in our previous works on automatic dialog act recognition.

Yet, the analysis of errors still suggests several ways to improve the quality of this initial corpus. First, extending the dialog act tag set may be useful, as 8% of extracted sentences can not be classified within the four basic dialog acts. Furthermore, additional rules shall also be included, as more than 70% of the errors have been considered by the human annotator as relatively easy to handle via lexical or grammatical rules. Finally, we plan to use semi-automatic methods, for instance based on the Expectation-Maximization algorithm or Active Learning, to reduce the number of remaining errors.

ACKNOWLEDGEMENTS

This work has been partly supported by the Ministry of Education, Youth and Sports of Czech republic grant (NPV II-2C06009).

REFERENCES

Austin, J. L. (1962). How to do Things with Words. *Clarendon Press, Oxford*.

Botha, G. and Barnard, E. (2005). Two approaches to gathering texte corpora from the world wide web. In *In Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa*, pages 194–197, Langebaan, South Africa.

Dhillon, R., S., B., Carvey, H., and E., S. (2004). Meeting Recorder Project: Dialog Act Labeling Guide. Technical Report TR-04-002, International Computer Science Institute.

Jekat *et al.*, S. (1995). Dialogue Acts in VERBMOBIL. In *Verbmobil Report 65*.

Jeong, M. and Lee, G. G. (2006). Jointly predicting dialog act and named entity for spoken language understanding. In *IEEE/ACL 2006 Workshop on Spoken Language Technology*.

Král, P., Cerisara, C., and Klečková, J. (2006). Automatic Dialog Acts Recognition based on Sentence Structure. In *ICASSP'06*, pages 61–64, Toulouse, France.

Král, P., Cerisara, C., and Klečková, J. (2007). Confidence Measures for Semi-automatic Labeling of Dialog Acts. In *ICASSP'07*, pages 153–156, Honolulu, Hawaii, USA.

Král, P., Pavelka, T., and Cerisara, C. (2008). Evaluation of Dialogue Act Recognition Approaches. In *MLSP'08*, pages 492–497, Cancun, Mexico.

Maeda, K., Lee, H., Medero, S., Medero, J., Parker, R., and Strassel, S. (2008). Annotation Tool Development for Large-Scale Corpus Creation Projects at the Linguistic Data Consortium. In *Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Marocco. http://www.lrec-conf.org/proceedings/lrec2008/.

Pavel, D. S. H., Sarkar, A. I., and Khan, M. (2006). A proposed automated extraction procedure of bangla text for corpus creation in unicode. In *In Proceedings of International Conference on Computer Processing de Bangla, ICCPB*, pages 157–161, Dhaka, Bangladesh.

Sarmento, L., Carvalho, P., Silva, M. J., and de Oliveira, E. (2009). Automatic creation of a reference corpus for political opinion mining in user-generated content. In *TSA '09: Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 29–36, New York, NY, USA. ACM.

Zhang, Y., Wu, K., Gao, J., and Vines, P. (2006). Automatic acquisition of chinese-english parallel corpus from the web. in: Ecir2006. In *In Proceedings of 28th European Conference on Information Retrieval*, pages 420–431. Springer.