

Annotation of Polish spoken dialogs in LUNA project

Agnieszka Mykowiecka*, Krzysztof Marasek†, Małgorzata Marciniak*,
Joanna Rابيةga-Wisniewska*, Ryszard Gubrynowicz†

*Institute of Computer Science, Polish Academy of Sciences
J. K. Ordona 21, 01-237 Warsaw, Poland
{Agnieszka.Mykowiecka, Malgorzata.Marciniak, jrابيةga}@ipipan.waw.pl

†Polish Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
{kmarasek, rgubryn}@pjwstk.edu.pl

Abstract

In this paper we present general assumptions and goals of the LUNA (spoken Language UNDERstanding in multilingual communication systems) project. We describe the process of collecting a Polish corpus of spoken dialogs and the accepted annotation schema of this corpus at several levels, from transcription of dialogs and morphosyntactic analysis, to semantic and dialog acts annotation.

1. Introduction

LUNA started in September 2006 as a STREP project in the 6th framework of the European Commission. Its main goal is to create a robust and effective spoken language understanding (SLU) module, which can be used to improve the speech-enabled telecom services in multilingual context. The consortium is built around two main actors: Loquendo — an Italian company specialized in speech products for telecom markets acting as a project coordinator and Prof. Renato de Mori from University of Avignon, who is a Technical Manager to the consortium. Remaining project partners involve both industry and academia: RWTH Aachen, University of Trento, France Telecom R&D (FT), CSI-Piemonte (CSI) and two Polish partners: Institute of Computer Science Polish Academy of Sciences and Polish-Japanese Institute of Information Technology.

1.1. Goals

The SLU tries to tackle the problem of human-machine spoken dialogs adding higher level models and descriptors to the hypotheses produced by a speech recognizer. It is expected that usage of semantic models will leverage the quality of dialog system as well as an acceptance of human-computer interaction.

LUNA project focuses on five main scientific problems:

- Language Modeling for Speech Understanding

Language Model (LM), a basic component of speech recognizer will be modified to cover both n-gram and semantic constituents, also novel clustered LMs will be used in order to characterize concepts (De Mori, 1998).

- Semantic Modeling for Speech Understanding

Methods of machine translation will be applied to the word lattice with POS tags added in order to extract semantic constituents from dialogs. These will

be used to create concept tags hypotheses or conceptual structures. To increase the likelihood of models, complex confidence measures will be built based on acoustic, linguistic and semantic indicators.

- Automatic Learning

To refine semantic models on-line learning methods and discriminative training will be used. Especially dynamic updating of knowledge sources may improve ASR and SLU results.

- Robustness issues for SLU

Previous studies indicate the necessity of new confidence indicators related to semantic interpretation. Moreover, out-of-vocabulary words highly biases the results of the ASR — SLU might help to fight against this problem.

- Multilingual portability of SLU components

Portability is one of the central goals of the LUNA. Portability will be regarded both from one language to another and from one application to another. The possibility of component re-uses and adaptation to a similar domain will be investigated.

The main results of the project will be speech dialogs corpora (human/human, human/machine or human/Wizard-of-Oz) for three well-defined domains for French, Italian and Polish. The data will be annotated on several levels up to semantic constituents and will be used to train and validate models and methods of SLU.

FT will enlarge and develop two corpora of human/machine dialogs concerning stock exchange and customer support service. This part of the project stress on the semantic annotation, taking advantage of the MEDIA project experience (Bonneau-Maynard and et al, 2005).

CSI and Trento University will collect 500 human/human dialogs in the customer service domain, transcribe them and annotate at several levels. Then, a corpus of 500 dialogs human/machine will be prepared.

Polish corpus will be described more precisely in the following sections.

2. Corpus of Polish spoken dialogs

Polish corpus consists of around 12 000 calls made to a transportation information center during late March to 1st half of May 2007. The corpus of dialogs was collected from the Warsaw city transportation information center known as ZTM Service Center which telephone number is posted on all transportation line stops in the city. It is a small contact center staffed daily by two to four persons who typically field 200-300 calls per day with an average call duration of 1-2 minutes. During data collection period one of the project members conducted on-site observations of call center agents to obtain a better understanding of the work-flow, group dynamics and interaction between call center agents. Within the LUNA project 500 dialogs will be annotated. The remaining data will be stored and available for the future analysis and processing.

The audio signal from the call center is monitored and recorded automatically on the PC (initial sampling frequency 32 kHz, 16 bits, resampled to 16 kHz) using a special device, which enables dual-channel recording of the telephone conversations.

Preparation of LUNA human-to-human dialogs corpus is underway. At the time of writing, speech data have been collected and substantial part of it has been transliterated and annotated on the acoustic level. This is a hard work of which main problems are the following:

- Noisy recordings: most of them are in adverse acoustic conditions (many calls are carried when driving or waiting on a street, thus a lot of external noises, wind noises, etc. disturb the speech).
- Low quality of speech transmission: many calls are over GSM when moving or using low quality microphones; overall low level of speech signal.
- Disfluencies, hesitations, repetitions, grammatical and pronunciation errors, mistakes and corrections, jargons, false starts and disruptions, word and phrase fragments, filled pauses, indeed all possible spontaneous speech effects are observed.
- Long pauses and long speeches, quite often not at all relevant to the main topic of the dialog.
- Strong emotions (quite often negative) of the speaker influencing her/his articulation.

The prevalent dialogs in the call center concern time schedules of trams, buses and metro lines. However, there are also other very frequent calls concerning the route paths between given points of the city (lines to be used, nearest stops, transfers, trip duration).

Before further processing, registered calls are classified according to a dialog topic. Some calls are irrelevant to the call-center competence and they are not further processed (about 10%). In many other dialogs there are still parts of various significance for identification of the calls topics and it is possible to mislead a subject of conversation.

3. Segmentation and transcription

An annotator is converting the recordings into plain texts applying Transcriber (Barras et al., 1998). Every

conversation is divided into turns referred to the caller and the operator, respectively. The transcription output is an XML file which includes the dialog text and some meta-data referring to articulation distortions, speaker and non-speaker noises, and time-stamps of the beginning and the end of each turn.

As the aim of this annotation level is to gain the natural, uninterpreted utterances, an annotator should transcribe manually segmented turns without modification, if possible. To make the further processing easier and the form of a transcribed file more uniform within the project, the following conventions have been adopted:

- It is agreed to capitalize words following the standards of each language, e. g. *Galeria Mokotów, Bitwy Warszawskiej*.
- Spellings are transcribed using capital letters separated by spaces and tagged with a symbol *pron=SPELLED*. The same approach is used to transcribe acronyms that are spelled, e. g. *[pron=SPELLED-] A W F [-pron=SPELLED]*.
- Acronyms pronounced as words are written in capitals without dots or spaces between letters, e. g. *PEKAES*.
- Foreign words or acronyms are transcribed in their original orthographic form and tagged with a symbol *lang=* and the name of the language, e. g. *[lang=English-] McDonald [-lang=English]*.
- If an acronym is spelled with foreign pronunciation, an annotator combines the tags *lang=* and *pron=SPELLED*, e. g. *[lang=English-] [pron=SPELLED-] C D [-pron=SPELLED] ROM [-lang=English]*.
- Numbers are transcribed as words following the standards of each language.
- Only the actually spoken part of a word is transcribed. Possible truncation is marked with *lex=~*, as annotators do not interpret a word or an utterance at this level.
- In case of mispronunciation, the correct form is transcribed with an indication that it has been mispronounced, e. g. *[pron=*-] placu [-pron=*] Narutowicza*.
- In general, the transcription does not include punctuation marks.
- Words that cannot be recognized, are transcribed with the symbol *pron=***.
- Tag *lex=FIL* is used to represent pause fillers, hesitations and articulatory noises as breath, laugh, cough, etc. Non human noises are annotated with the tag *noise*. Silence is annotated only if it lasts more than 1 second – as *sil*.

A sample of transcribed conversation concerning the user's question referring to city communication network is presented in Fig. 1.

user: [lex=FIL] dobry wieczór jak mogę się dostać z [pron=*-] placu [-pron=*] Narutowicza do Galerii Mokotów
user: good evening how can I get from Narutowicz square to Mokotów Gallery
operator: proszę podjechać tramwajem do Bitwy Warszawskiej i stamtąd sto osiemdziesiąt sześć
operator: you should go by tram to Warsaw Battle (first) and then (with) one hundred six

Figure 1: Sample of the transcribed text

4. Morphosyntactic annotation

After transcription, the set of dialogs will be annotated with the morphosyntactic tags. The annotation will be done in two steps.

First, it is agreed to attach the POS tags to each word and to identify its morphological characteristics. As the project concerns three different languages, the partners have adopted the recommendations of EAGLES (<http://www.ilc.cnr.it/EAGLES96/>) for the morphosyntactic annotation of text corpora and computational lexicons and have defined for each language a core set of tags consistent with international standards. The result of the morphological analysis will be stored in an XML file in format presented in Fig. 2.

```

jak mogę się dostać z placu Narutowicza do Galerii
Mokotów?
how can I get from Narutowicz square to Mokotów Gallery?
<words>
<w id="1" word="jak" lemma="jak" POS="PINT"
morph="-" />
<w id="2" word="mogę" lemma="móc" POS="VA"
morph="1.sg.pres.ind.imperf" />
<w id="3" word="się" lemma="się" POS="PART"
morph="-" />
<w id="4" word="dostać" lemma="dostać" POS="VV"
morph="inf.perf" />
<w id="5" word="z" lemma="z" POS="PreP"
morph="-" />
<w id="6" word="placu" lemma="plac" POS="Nc"
morph="m3.gen.sg" />
<w id="7" word="Narutowicza" lemma="Narutowicz"
POS="Np" morph="m1.gen.sg" />
<w id="8" word="do" lemma="do" POS="PreP"
morph="-" />
<w id="9" word="Galerii" lemma="Galeria" POS="Np"
morph="fem.gen.sg" />
<w id="10" word="Mokotów" lemma="Mokotów"
POS="Np" morph="m3.nom.sg" />
</words>

```

Figure 2: Example of morphological annotation

Second, based on the morphological annotation the text of the dialogs should be segmented into elementary syntactic chunks. The aim of syntactic description is to group the words into basic nominal phrases and verbal groups. At this level the partners also follow the rec-

ommendation for syntactic annotation of corpora of EAGLES. The syntactic segmentation is shown in Fig. 3.

```

jak mogę się dostać z placu Narutowicza do Galerii
Mokotów?
how can I get from Narutowicz square to Mokotów Gallery?
<chunks>
<chunk id="1" span="word_1" cat="PINT" />
<chunk id="2" span="word_2..word_4" cat="VP" />
<chunk id="3" span="word_5" cat="Prep" />
<chunk id="4" span="word_6..word_7" cat="NP" />
<chunk id="5" span="word_8" cat="PreP" />
<chunk id="6" span="word_9..word_10" cat="NP" />
</chunks>

```

Figure 3: Example of syntactic annotation

5. Semantic annotation

5.1. Representation of semantics — domain ontology

A widely accepted method of representing semantics is defining domain ontologies. Although their universality can be questioned and the reuse of resources is not easy, see (Paslaru-Bontas, 2007), it is still the best existing solution for achieving portability. For the purpose of describing Warsaw public transportation, we decided to develop a new ontology using OWL (DL sublanguage) standard. Our *CityTransport* ontology covers:

- Different transportation means (buses, trams, local trains and metro), their routes, stops and timetables;
- Town topology in the aspects needed for traveling (stops, important buildings, streets' names);
- Trips plans using one or more transport means.

The ontology describes a typology of classes. Fig. 4 shows a fragment of a tree of classes which includes among others description of places in a town. Apart from the typology, an ontology allows for describing class properties. We can define types of values and cardinality. For example, for every identified place we define its name, address and a set of public transportation stops which are nearby (see Fig. 5).

```

TownLocation
  BuildingOrOtherPlace:
    Building, Park, Cemetery, ...
  TownSection
    District, StreetOrSquare
  PublicTransportPlace
    PublicTransportStop, MetroStation, TrainStation
  PublicTransportLine:
    BusLine, TramLine, TrainLine, MetroLine
  Trip
  Passage

```

Figure 4: Fragment of the ontology

```

BuildingOrOtherPlace:
  Named (some values from TownPlacesNames)
  CloseTo (multiple PublicTransportStop)
  LocatedAt (one value from StreetName)
BusLine
  HasBusName (one value from BusNames)
  HasTimetable (one value from Timetable)
  HasRoute (several values from BusRoute)
Passage
  HasStartStop (one value from PublicTransportPlace)
  HasEndStop (one value from PublicTransportPlace)
  DoneBy (one value from PublicTransportLine)

```

Figure 5: Properties of selected classes

5.2. Attribute level annotation

The first semantic level of annotation concerns assigning attributes' names to phrases which realize them. Names and values of the attributes are taken from the defined transportation ontology. Fig. 6 shows attribute annotation for the example given in Fig. 3. In this sentence, three objects described in the ontology are identified: a trip taking action, a name of a building, a street or a square.

```

<concept id=c_1 span="word_1..word_5"
  attribute="action" value="TripPlanRequest" />
<concept id=c_2 span="word_6..word_7"
  attribute="StreetOrSquare" value="plac Narutowicza" />
<concept id=c_3 span="word_9..word_10"
  attribute="Building" value="Galeria Mokotów" />

```

Figure 6: Attribute level annotation example

5.3. Predicate level annotation

Predicate structure is represented in a FrameNet like format as a connection between phrases and roles which they realize, (Fillmore, 1968; Baker et al., 2003). As there is no Polish FrameNet, from the collected dialog corpus we will extract all domain related verbs and define all needed frame patterns. Names of roles are taken from our domain ontology. A tag `link_concept` is used to connect the frame level roles to concepts of the attribute level, see Fig. 7.

It may happen that a dialog fragment does not contain the explicit identification of a frame (e. g. instead of: *a do Galerii Mokotów?*, is – *a do Galerii?* ‘*and to Gallery Mokotów?*’ – ‘*and to Gallery?*’) In such a case, an annotator has to try to infer the appropriate frame. If it is not possible, the frame should be annotated as *unknown*. An additional label *other* will be used to annotate frames which are not in the predefined frame set.

5.4. Anaphoric relations

Annotation of anaphoric relations in LUNA is based on the ARRAU (AnaphoRa Resolution And Underspecification) project (Artstein and Poesio, 2006),

```

Frame: TPR (TripPlanRequest)
Frame-elements: TripStart, TripEnd, LineName
set3={fe_1, fe_2, fe_3}
<fe id="fe_1" span="word_1..word_5" frame="TPR"
  frame-element="target" link_concept="c_1"
  member="set_3" />
<fe id="fe_2" span="word_5..word_7" frame="TPR"
  frame-element="TripStart" link_concept="c_2"
  member="set_3" pointer="fe_1" />
<fe id="fe_3" span="word_8..word_10" frame="TPR"
  frame-element="TripEnd" link_concept="c_3"
  member="set_3" pointer="fe_1" />

```

Figure 7: Predicate level annotation example

but is restricted to NPs (including pronouns and adverbial temporal expressions) referring to domain concepts.

An NP is marked with a label *new* or *given* depending on whether the related to it concept appears in a dialog for the first time, or it refers to a previously mentioned concept. If the status of a phrase is marked as *given*, a set of its antecedents has to be established (labels *single-phrase-antecedent* or *multiple-phrase-antecedent*). A phrase may refer only to one or to all elements of an antecedents set. The first case is indicated by a label *ambiguity*. A label *link_concept* connects the coreference level with the domain concept level (see Fig. 8).

Moreover, on this level of annotation we may express some relations between concepts assigned to phrases. For example, if a question concerns ‘public transport’ and an answer contains an information about ‘a bus’ we indicate the relation *SubClassOf* between the appropriate concepts.

```

jak mogę się dostać z / placu Narutowicza / do / Galerii
Mokotów /
how can I get from / Narutowicz square / to / Mokotów
Gallery /
< coref id="e_1" span="word_6..word_7" inf_status="new"
  related="no" link_concept="c_1" /> ...
proszę podjechać /tramwajem / do /Bitwy Warszawskiej / i
/ stamtąd / sto osiemdziesiąt sześć /
you should go / by tram / to / Warsaw Battle / (first) and
/ then / (with) one hundred six /
...
< coref id="e_4" span="word_15..word_16"
  inf_status="new" related="no" link_concept="c_4" />
< coref id="e_5" span="word_18" inf_status="given"
  single_phrase_antecedent="e_3" /> ...

```

Figure 8: Coreference level annotation example

6. Dialog acts

Dialog acts are the last level of annotation in the LUNA project and is not obligatory. A dialog model created from dialog acts is necessary in case of automatic dialog system creation, and can be useful in automatic understanding of speech for solving ambiguities. The dialog act

annotation in LUNA is based on the DAMSL project — Dialogue Act Markup in Several Layers (Core and Allen, 1997).

The following dialog acts have been chosen in LUNA as the starting point. The list of dialog acts may be extended for different domains by the project partners according to their needs:

- Forward looking function:
 - Statement;
 - Action-directive / Open option;
 - Committing-speaker-future-action;
 - Info-request.
- Backward looking function:
 - Answer;
 - Accept / Reject;
 - Signal-understanding / Signal-non-understanding.

In the project, it is assumed that the basic unit attributed with dialog acts is a fragment of a turn, and that more than one dialog act can be assigned to it. For example the first sentence from the dialog in Fig. 8 should be marked as *Info-request* and the second one should be marked as *Answer* to the previous question and *Committing-speaker-future-action* because it suggests the addressee to perform some action.

The final set of dialog acts for Polish will be chosen after careful analysis of a greater number of dialogs. We have proposed already to return to some ideas from the DAMSL project and to annotate phrases with: *Opening*, *Closing* and add *Politeness* or *Impoliteness* tags. It would be also useful to connect a question with its answer or an offer with its acceptance or rejection. Moreover, we suggest to allow marking of these dialog parts which are not interesting from our point of view as *Irrelevant*.

7. Summary

The result of the project will be the first corpus of Polish spoken dialogs annotated with morphological, syntactic and semantic information. So far, only a few Polish speech corpora have been collected at all. One of the first research done on speech data was undertaken in the seventies by K. Pisarkowa and concerned Polish syntax of a telephone conversation (Pisarkowa, 1975). Although the linguist had recorded the dialogs for the study, that corpus is not available anymore. SpeechDAT Polish (Heuvel et al., 2001) is the only widely distributed Polish speech database collected over telephone lines, but it does not contain dialogs nor spontaneous speech.

The collected corpus will be publicly available for research purposes and it can be used to test various linguistic and application oriented hypotheses. Especially effects of human-human interaction and spontaneous speech phenomena are worth of more detailed investigation.

8. Acknowledgements

This work is supported by LUNA (IST 033549) project. The authors would like to thank Warsaw Transport Authority (ZTM Warszawa) and its call center employees for their support and cooperation.

9. References

- Artstein, R. and R. Poesio, 2006. Arrau annotation manual (trains dialogues). Technical report, University of Essex.
- Baker, Colin F., Charles J. Fillmore, and Beau Cronin, 2003. The structure of the framenet database. *International Journal of Lexicography*, 16:281–296.
- Barras, C., E. Geoffrois, Z. Wu, and M. Liberman, 1998. Transcriber: a free tool for segmenting, labeling and transcribing speech. In *First International Conference on Language Resources and Evaluation (LREC)*.
- Bonneau-Maynard, H. and S. Rosset et al, 2005. Semantic annotation of the media corpus for spoken dialog. In *ISCA Interspeech*, volume ISCA Interspeech. Lisbon.
- Core, Mark G. and James F. Allen, 1997. Coding dialogues with the DAMSL annotation scheme. In David Traum (ed.), *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*. Menlo Park, California: American Association for Artificial Intelligence.
- De Mori, R., 1998. *Spoken dialogues with computers*. Academic Press.
- Fillmore, Charles J., 1968. The case for case. In E. Bach and R. Harms (eds.), *Universals in Linguistic Theory*. New York: Holt, Rinehart and Winston.
- Heuvel, H., J. Boudy, and Z. Bakcsi et al, 2001. Speechdat-e: Five eastern european speech databases for voice-operated teleservices completed. In Paul Dalsgaard et al. (ed.), *Eurospeech 2001 Scandinavia, 7th European Conference on Speech Communication and Technology*. Aalborg, Denmark.
- Paslaru-Bontas, E., 2007. *A Contextual Approach to Ontology Reuse Methodology, Methods and Tools for the Semantic Web*. Ph.D. thesis, Fachbereich Mathematik u. Informatik, Freie Universität Berlin.
- Pisarkowa, K., 1975. *Składnia rozmowy telefonicznej*. Wydawnictwo PAN.