Unsupervised Multi-Topic Labeling for Spoken Utterances

Sebastian Weigelt, Jan Keim, Tobias Hey, Walter F. Tichy
Karlsruhe Institute of Technology
Institute for Program Structures and Data Organization
Karlsruhe, Germany
weigelt@kit.edu, jan.keim@kit.edu, hey@kit.edu, tichy@kit.edu

***Abstract*—Systems such as Alexa, Cortana, and Siri appear rather smart. However, they only react to predefined wordings and do not actually grasp the user's intent. To overcome this limitation, a system must grasp the topics the user is talking about. Therefore, we apply unsupervised multi-topic labeling to spoken utterances. Although topic labeling is a well-studied task on textual documents, its potential for spoken input is almost unexplored. Our approach for topic labeling is tailored to spoken utterances; it copes with short and ungrammatical input.**

The approach is two-tiered. First, we disambiguate word senses. We utilize Wikipedia as pre-labeled corpus to train a naïve-bayes classifier. Second, we build topic graphs based on DBpedia relations. We use two strategies to determine central terms in the graphs, i.e. the shared topics. One focuses on the dominant senses in the utterance and the other covers as many distinct senses as possible. Our approach creates multiple distinct topics per utterance and ranks results.

The evaluation shows that the approach is feasible; the word sense disambiguation achieves a recall of 0.799. Concerning topic labeling, in a user study subjects assessed that in 90.9% of the cases at least one proposed topic label among the first four is a good fit. With regard to precision, the subjects judged that 77.2% of the top ranked labels are a good fit or good but somewhat too broad (Fleiss' kappa κ = 0.27).

I. Introduction

Conversational interfaces (CI) are a recent trend in human computer interaction. Today, millions of users communicate with virtual assistants such as Alexa, Cortana, or Siri. However, such systems often struggle to actually grasp the user's intent. Although they appear rather smart, Alexa and the like merely react to predefined commands. Users will soon expect such systems to understand increasingly complex requests. Thus, techniques for (deep) spoken language understanding (SLU) are needed. We propose to apply topic labeling to spoken utterances as one building block of a comprehensive intent model. Topic modeling and labeling has already proved useful on textual documents; it has been applied to many tasks, such as text summarization, machine translation, and sentiment analysis [1]. However, topic labeling has rarely been adapted to spoken utterances scenarios [2]. Most likely this is the consequence of differing boundary conditions. Spoken language is typically ungrammatical. Thus, common techniques for natural language (pre-)processing (NLP) cannot be applied. Furthermore, utterances – be it dialog acts, virtual assistant interactions, or instructions for household robots – are short in comparison to text documents. This limits the usefulness of contextual information to a minimum. An exemplary input of that kind might be, "Hey robot, take – uhm – the apple – err – the orange from the fridge." Even though the utterance is rather short, it encompasses three topics: *Domestic Robotics*, *Fruits*, and *Home Appliances*. Present approaches for topic labeling cannot cope with such conditions as they either rely on NLP or contextual models.

Our approach is influenced by a number of related approaches to topic labeling on documents. However, it is customized to the challenges of short, spoken utterances. Our approach is two-tiered. First, we perform word sense disambiguation (WSD). We have adapted the approach by Mihalcea [3] and Mihalcea and Csomai [4]. The approach uses Wikipedia as a pre-labeled corpus and applies a naïve-bayes classifier. Nouns are labeled with Wikipedia articles. Second, we use the word sense labels to determine topic labels. To this end, we build so-called sense graphs. Beginning with the Wikipedia articles attached to nouns in the utterance, we use relations in DBpedia to construct graphs. Afterwards, we determine the *most central terms*, which we take to be the topic labels for the utterance. We have implemented two different strategies for graph centrality. The first generates topic labels for dominant terms, i.e. the most frequent senses, in the utterance. The latter covers all terms. Both produce multiple labels for each utterance. The labels carry confidences, which we derive from the graph centrality value. The contribution of the paper is two-fold:

1) An adaptation of the WSD approach by Mihalcea and Csomai to short utterances, including an evaluation on a Wikipedia data set plus an additional evaluation on a corpus for programming in spoken language.
2) An implementation and evaluation of unsupervised multi-topic labeling tailored to short, spoken utterances.

The remainder of the paper is structured as follows: First, we discuss related work in Section II. In Section III we introduce our approach for unsupervised multi-topic labeling and evaluate it in Section IV. Finally, we discuss areas of application (Section V) before we conclude the paper in Section VI.

II. Related Work

Topic labeling is typically preceded by a topic modeling step that determines sets of terms that are supposed to share

the same topic. Afterwards, meaningful labels are assigned to these topics. Many approaches rely on the so-called *Latent Dirichlet Allocation* (*LDA*) introduced by Blei et al. [5] to create a topic model [6]–[10]. *LDA* is a generative probabilistic model for collections of discrete data such as text documents. It uses word distributions across a given set of documents to derive topics from word occurrences. Hence, an *LDA* topic model comprises a fixed number of topics that consist of words which often occur together.

To determine meaningful labels, some approaches derive labels directly from the given text [8], [11], assuming that a label can be found within the given text. However, this assumption may not hold. Often, a document does not contain appropriate labels; i.e. for certain topics no abstract term is ever mentioned. Additionally, text-based approaches usually suffer from challenges such as synonyms or spelling errors. Thus, advanced approaches incorporate additional information to gain a deeper understanding of a topic. Usually, these approaches map words that represent a topic to knowledge databases. Then, they create graph or tree structures based on relations in the knowledge database (e.g. Magatti et al. [9]). Another approach of that kind was introduced by Hulpus et al. [10]. The authors calculate a topic model with *LDA* and then determine central nodes in a so-called topic graph, which they build from DBpedia concepts and relations. Central concepts form the topic labels. All above-mentioned approaches use *LDA* to some extent, which is a statistical model. Therefore, its performance depends on the available amount of data. As spoken utterances are rather short, *LDA* does not produce reliable results. Hence, *LDA*-based approaches are infeasible in our context.

Some related approaches do not rely on *LDA*. Coursey et al. [12] create graphs based on Wikipedia articles (nodes) and the proximity of the containing words (edges). They determine central nodes with the help of a biased PageRank algorithm and use the article names of these nodes as topic labels. Aker et al. [13] use the Markov Clustering Algorithm for topic modeling. Allahyari and Kochut [14] adapt *LDA*; they introduce a latent variable called *concept*. The concepts are DBpedia concepts and are used to build graphs. Recently, combined approaches are used; they either join different topic labeling approaches (e.g. Gourru et al. [15]), or incorporate concepts from other research areas, e.g. word embeddings [16]. However, they also require long documents to unfold their potential. Thus, they are inappropriate for short, spoken utterances.

In the field of SLU various approaches use phonetic information to model topics. Cerisara [17] creates a semantic lexicon from phonetic information and creates topic models by hierarchical clustering. Hazen et al. [18] and Siu et al. [19] propose similar approaches to model topics. However, none of the approaches actually label topics.

In summary, all above are inapplicable to determine topic labels for short utterances. *LDA*-based approaches (and others intended for texts) require long documents and present SLU approaches only model topics but do not determine labels.

III. Approach

Our approach for unsupervised multi-topic labeling is inspired by topic modeling and labeling approaches for text documents. However, it does not rely on a generative probabilistic model such as *LDA*. This is mainly because *LDA* is not applicable on short documents. Additionally, *LDA* can only distinguish a fixed number of topics. However, in our context the number of topics is uncertain in advance. Unlike *LDA*-based approaches, we build topic graphs for the entire input, i.e. each spoken utterance. We use data from DBpedia to create these graphs; articles are nodes and relations form edges. We use a biased PageRank algorithm to determine multiple central articles per utterance, which we use as topic labels. Therefore, we are relieved from the challenge of creating meaningful labels. Instead, we only have to determine which term is the most fitting for a topic. The approach requires word sense labels as starting point for the construction of sense graphs. For this, we adapt the approach by Mihalcea and Csomai [4] that uses Wikipedia as a pre-labeled corpus for WSD. It uses naïve-bayes classification to attach Wikipedia articles (as senses) to nouns.

In Subsection III-A we present our adapted re-implementation of their WSD method. Afterwards, we describe our unsupervised multi-topic labeling approach for spoken utterances in detail in Subsection III-B.

A. Word Sense Disambiguation

Supervised classification tasks require manually attached labels for training, which is time-consuming and costly. Additionally, in the case of word sense disambiguation human annotators often disagree. Mihalcea and Csomai tackle this issue by using Wikipedia as a pre-labeled corpus for word senses. The basic idea is as follows. Relevant terms (mostly nouns) in a Wikipedia article each have a link attached to the respective explanatory article. Thus, links can serve as manually annotated word senses.

Links are added by the article's authors (most commonly), who are supposed to be domain experts. Therefore, Mihalcea and Csomai assume that the links are correct. Also, Wikipedia is growing steadily and the quality of articles improves over time through continuous inspection by the community. Even though the latter is arguable, the quality of Wikipedia articles surely has improved since Mihalcea and Csomai first implemented their approach in 2007. Further details about the original approach may be found in Mihalcea [3] and Mihalcea and Csomai [4].

We adopt the idea to use Wikipedia as a pre-labeled corpus for word senses. However, we altered the classification process slightly. We also use a naïve-bayes classifier and similar features: the ambiguous word, its part-of-speech (POS) tag, the three words to the left and right of the ambiguous word, and their POS tags, as well as the first nouns and verbs to the left and right. To increase the impact of the ambiguous word over its contextual features we weighted it tenfold (contrary to Mihalcea and Csomai, who did not alter weights). We filter out stop-words. Mihalcea and Csomai additionally use so-called

TABLE I: DBpedia relations used to build sense graphs.

relation	relates a concept to
dcterms:subject	its Wikipedia category
skos:broader	less specific concepts
skos:narrower	more specific concepts
purlg:hypernym	superordinate concepts
purlg:meronym	concepts that form parts
purlg:synonym	synonymous concepts
rdfs:type	its DBpedia ontology entity
rdfs:subClassOf	its subclasses in DBpedia
rdfs:seeAlso	related concepts

context words, which are simply the most frequent words of the paragraph in which the ambiguous word appears. Context words are not feasible in our context, because short, spoken utterances do not consist of paragraphs. Even if we define a full utterance as paragraph, it is rather short with barely multiple occurrences of words that are not stop-words. We also skip the disambiguation of named entities. As they are usually unambiguous, there is no need to disambiguate these terms and their mere number impairs our classification model.

To train the classifier, we use a Wikipedia dump from August 2017. We prepare the data like Mihalcea and Csomai. We remove disambiguation pages, as they do not contain full sentences. For the same reason, we ignore info boxes and lists. Additionally, lists rarely contain links, which makes them useless. The same applies to quotes. We also remove links that lead to an article that considers a named entity; those are simply unusable topic labels.

We extracted 5,188,470 training instances. Among them are 283,173 different senses, of which 136,964 are unique. Unique senses are senses that are mentioned in one instance only. These unique senses account for 2.64% of the instances and 48.37% of the senses.

We can then use the trained model as a WSD classifier. Note that the classifier can only disambiguate nouns. However, contrary to Mihalcea and Csomai, our classifier attaches a label – i.e an Wikipedia article – to *all* nouns in the input.

B. Topic Labeling

Our approach for unsupervised multi-topic labeling is inspired by the sense graph idea proposed by Hulpus et al. [10]. However, they used *LDA* to determine topic models. As discussed before, we cannot use a statistical model such as *LDA* on short, spoken utterances. Instead, we directly determine topic labels and perform topic modeling implicitly. We assume that all nouns in the input are related at first (we discard that assumption later). We build sense graphs beginning with the word sense for each noun; we call these senses *initial senses*. We think of DBpedia as a graph with concepts (i.e. articles) as nodes and relations as edges and extract subgraphs. We traverse all chains of relations up to a distance of two to create the sense graphs[1]. The relations we use to build the sense graphs are listed in Table I.

Finally, we merge all sense graphs; the result is a topic graph. Hulpus et al. remove all disconnected subgraphs and proceed with the main graph only. Instead, we proceed with the entire graph, including all disconnected subgraphs. As each subgraph originates from different sense graphs, we assume that subgraphs represent different topic areas. Thus, subgraphs compensate the missing topic modeling step in our approach. Then, we can determine topics for each of the subgraphs, i.e. topic areas.

We continue with determining the central nodes of the graph. Hulpus et al. discuss different algorithms to determine graph centrality. However, none of them can cope with disconnected graphs. Instead, we apply a biased PageRank algorithm ([12], [20]); it gives more weight to nodes that correspond to the initial senses. The biased PageRank, i.e. the score $\mathrm{S}(V_i)$, is calculated as follows:

$$\mathrm{S}(V_i) = (1-d) * \mathrm{B}(V_i) + d * \sum_{j \in \mathrm{I}(V_i)} \frac{\mathrm{S}(V_j)}{|\,\mathrm{O}(V_j)|} \tag{1}$$

where $\mathrm{I}(V_i)$ is the set of incoming edges of node i and $\mathrm{O}(V_i)$ is the set of outgoing edges of node i. The constant d is the *damping factor*; in our implementation, we leave the default value unchanged ($d = 0.85$). The bias $\mathrm{B}(V_i)$ is defined as:

$$\mathrm{B}(V_i) = \frac{f(V_i)}{\sum_{j \in \mathrm{InitNodes}} f(V_j)} \tag{2}$$

where $\mathrm{InitNodes}$ is the set of nodes that correspond to the initial senses. Coursey et al. state, that $f(V_i)$ may vary regarding complexity and can be chosen freely. They also discuss different options for the choice of $f(V_i)$. For their approach they choose an $f(V_i)$ that was determined by a so-called *keyphraseness score*. However, this score requires a set of documents. In our context we consider only one utterance at once. Thus, we determine $\mathrm{B}(V_i)$ differently; we simply set $f(V_i)$ to 1 if V_i is a member of the initial node set:

$$\mathrm{B}(V_i) = \begin{cases} 0 & , V_i \notin \text{InitNodes} \\ \frac{1}{|\text{InitNodes}|} & , V_i \in \text{InitNodes} \end{cases} \tag{3}$$

Finally, we select the nodes from the topic graphs that will serve as topic labels. The selection of labels is contingent on the number of labels we create per utterance. On the one hand, with an increasing number of labels we observed that labels get too broad. Thus, the precision of our approach decreases. On the other hand, if we create too few labels, some senses from the utterance are not represented. As a consequence, the coverage decreases. We found that a good rule of thumb is to create twice as many labels as there are distinct senses in the input. This number of labels allows us to discover appropriate labels even for utterances with many distinct topics. At the same time, there is still a sufficient selection of labels for small inputs with only two or three senses.

To select labels our approach is configured with one of two strategies. The *top strategy* selects nodes as labels that are strongly connected to the original sense nodes. Therefore, we

[1]Ideally, the distance is as short as possible to generate meaningful sense graphs. Longer distances introduce an increasing semantic drift. However, if we traverse only one relation, less connected graphs are constructed, i.e. we might be unable to discover shared senses. Therefore, we follow the choice of Hulpus et al. [10] and use two as the distance value.

count the number of sense graphs in which each particular node occurs. We call this value *connectivity*. As the connectivity might be equal for multiple nodes, we use PageRank as second criterion. Therefore, the *top strategy* first selects the node with the highest connectivity. If there is a draw, the node with the higher PageRank is chosen. We repeat this procedure until the maximum number of labels is reached. However, we found that some parts of the utterances are inadequately represented. If a topic is mentioned with a few words only, it gets dominated by other topics. Therefore, we implemented a second strategy: the *max strategy*. This strategy determines the first label in the same manner as the *top strategy*. However, it then examines the set of senses from the initial input. If not all senses are represented through a label yet, it selects the node that covers the highest number of previously unrepresented senses. As before, if there is a draw, the node with the higher PageRank is chosen. If all senses are covered, the strategy continues as the *top strategy*. Again, the procedure is repeated until the maximum number of topic labels is reached. We use the PageRank values as confidences to rank the labels.

C. Example

In order to illustrate our approach, we discuss the exemplary utterance, "take the orange from the fridge and close the dishwasher." For the sake of simplicity we configure our approach to create a maximum of two labels. The first step is the disambiguation of nouns. For *orange* the WSD model picks the sense *Orange(fruit)* rather than other possible senses such as *Orange(color)* or *Orange(word)*. The approach determines *Refrigerator* and *Dishwasher* as senses for the two remaining nouns. Then, our approach creates a sense graph for each sense. The resulting sense graphs (SG_1, SG_2, and SG_3) are depicted in the upper half of Figure 1. Graphs may share sense nodes. In the example, the sense graphs SG_1 and SG_3 share – among others – the sense nodes *home* and *home appliance*. All sense graphs are merged at these shared nodes. The result is a topic graph that may consist of disconnected subgraphs. In the example, we have two topic subgraphs (TG_1 and TG_2) after merging. Connected sense graphs indicate that senses are topically similar. With the help of the topic graph we determine the labels. We obtain different results depending on the selected strategy. If we use the *top strategy*, graph connectivity is the key aspect. Thus, only sense nodes from the topic subgraph TG_1 are considered. From the set of sense nodes that connect the two sense graphs (highlighted in green) our approach selects the nodes with the highest PageRank (i.e. the yellow nodes). Thus, the *top strategy* creates the labels *home* and *home appliance*. However, the sense *Orange(fruit)* is not covered by these topic labels. Here, the result of the *max strategy* differs. First, it selects a node from those with the highest connectivity, too. Thus, the label *home* is also selected. However, the next label is drawn from the topic subgraph with senses that were not represented previously, here TG_2. Again, the node from the candidate set with the highest PageRank is chosen, here *fruit*. Thus, the *max strategy* selects the labels *home* and *fruit* (i.e. the red nodes).

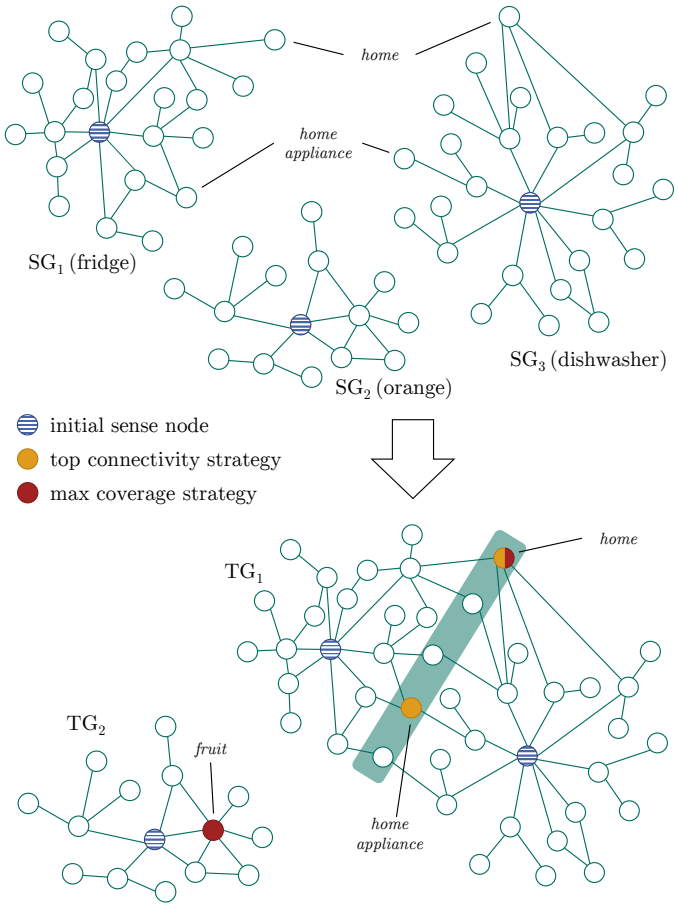

Fig. 1: The sense ($\mathrm{SG_i}$) and topic graphs ($\mathrm{TG_i}$) for the utterance, "take the orange from the fridge and close the dishwasher afterwards." The initial sense nodes are depicted blue-striped. The nodes selected as topic label by the *top strategy* are orange and the nodes selected by *max strategy* are red. The set of candidate nodes with equal connectivity in TG_1 are highlighted in green.

IV. Evaluation

To evaluate our approach we first assess the quality of the word sense disambiguation; its performance directly affects our unsupervised multi-topic labeling approach. As data sets we use Wikipedia and a speech corpus. The latter consists of 168 voice recordings from different user studies, gathered from 65 subjects. The subjects are between 18 and 50 years old, 21 are female and 44 male. Most of them are undergraduate and graduate students. All are non-native English speakers. However, their (self-assessed) English level is *advanced* on average. All recordings are instruction sequences for a household robot in eight different scenarios such as doing the laundry or preparing an instant meal. The recordings vary in length from 5 up to 80 seconds and in instructions from 2 up to 22. All recordings were manually transcribed according to the guideline by Kiesling et al. [21].

TABLE II: Results of the WSD evaluation.

Wikipedia	Speech Corpus		
avg. recall	precision	recall	F_1
.799	.894	.876	.885

We also evaluate our topic labeling approach on this corpus. To broaden the range of topics, we added synthetic utterances from other domains. We conducted a user study, where subjects manually evaluated the quality of the topic labels.

A. Word Sense Disambiguation

We first evaluate WSD on Wikipedia. We performed a customized ten-fold cross-validation. For each of the ten runs we drew 10,000 instances at random for testing; the remaining were used for training. Note that a full-blown ten-fold cross-validation with over five million instances is infeasible in our context. We determine the correctly predicted (true positives) and incorrectly predicted senses. As the number of instances is known in advance and our classifier predicts labels for all instances, every incorrectly predicted sense accounts for a false positive *and* a false negative. Therefore, precision and recall are the same here[2]. Mihalcea and Csomai distinguished false negatives and false positives in their evaluation. Their approach does not predict instances with a previously unseen surface form (during training phase). Thus, they removed all of these instances from the set of false positives. Consequently, we can compare recall only. Mihalcea and Csomai evaluated on a set of 85 Wikipedia articles drawn at random, which contained 7286 instances; we evaluated on ten times 10,000 random instances (see above). The results for our approach shown in Table II are encouraging. We achieve a recall/precision of 0.799. Despite the adaptations of the original approach (see Subsection III-A) and a test set differing largely in content and extent, this is comparable to the recall of 0.831 Mihalcea and Csomai reported.

In a second evaluation we used the speech corpus. Here, we prepared a gold standard for each noun. The manual transcriptions of the 168 recordings contained 1060 nouns in total. Note that in this evaluation we do not know instances in advance. To obtain the instances we have to identify nouns (except named entities) with a POS tagger. Again, false positives and false negatives encompass all incorrect labels. Additionally, all missed instances are false negatives. The results (shown in Table II) are promising. We expected a drop in classification quality, as the task is more complex (additionally determine instances) and the domain is different from the training set. Instead, our approach achieves a recall of 0.876 and a precision of 0.894 (F_1 0.885). 21 instances were not disambiguated due to incorrect POS tags produced by our POS tagger. Some incorrectly classified senses are due to nouns that have no corresponding article on Wikipedia, e.g. the word "front". There is no Wikipedia article describing the concept of *the side that is forward or prominent*. In such cases our approach retrieves incorrect senses.

Nevertheless, our results are encouraging. They show that the approach is feasible, even for domains where the content differs largely from Wikipedia articles (ungrammatical sequences mostly uttered in imperative mood vs. descriptive texts). Thus, the approach proves highly advisable in all contexts, where training of a custom WSD classifier is impossible because of data sparseness (as in our case) or too expensive.

[2]It is more common to use *accuracy* in this case, which is calculated equally. We kept the notions *precision* and *recall* for comparability with Mihalcea and Csomai

TABLE III: Distribution of the assessed quality of the top-k ranked topics produced with *max* and *top strategy*.

	good fit		too broad		inconv.		unrelated	
k	max	top	max	top	max	top	max	top
1	**.530**	**.530**	**.242**	**.242**	**.045**	**.045**	**.182**	**.182**
2	**.447**	.424	.167	**.182**	**.106**	**.106**	**.280**	.288
3	**.449**	.444	.141	**.152**	**.157**	.146	**.253**	.258
4	**.432**	.420	**.144**	.140	.155	**.167**	**.269**	.273
5	**.381**	.369	**.145**	.136	.176	**.182**	**.298**	.312
all	**.368**	.340	**.138**	.132	.179	**.200**	**.315**	.329

B. Topic Labeling

Evaluating the quality of a topic labeling approach is demanding; one cannot easily provide a gold standard. Usually it is unclear what the correct label is and if it is the only one fitting. Therefore, we performed a user study; it is similar to the study conducted by Hulpus et al. [10]. Six subjects participated in this study; all were graduate students from different faculties, four male and two female, aged 22 to 27. We drew 16 recordings from the speech corpus at random and provided manual transcripts for each. Additionally, we created six synthetic utterance transcriptions. They are comparable to the corpus recordings in regard to linguistic complexity and length. However, these transcripts are from other domains: drone control, child's playroom, and virtual assistants. Thus, we can evaluate our approach on a broader range of domains. We used our WSD classifier trained on the entire Wikipedia dump to label each noun. Based on the sense labels we created topic labels for each utterance with the *max strategy* and the *top strategy*. We presented the utterances to the subjects together with the topic labels. The labels are ordered according to their confidence values. The total number of topics per utterance varies from four to ten. The subjects were asked to rate each label either as *good fit*, *related but too broad*, *related but inconvenient* or *unrelated*. We divided the subjects into two groups that assessed the labels of eleven utterances each. Thus, all labels were assessed by three annotators. We use Fleiss' Kappa (κ) to measure the inter-annotator agreement; the determined κ value is 0.27. According to Landis and Koch, this indicates a fair agreement [22]. Hulpus et al. reported a similar κ-value. This outcome illustrates that, although topic labels are quite subjective, shared preferences between annotators are present.

The assessment results are depicted in Table III. It shows the distribution for the top-k topic labels and for all labels (the best results per rank and category are printed in bold). The *good fit*-labels can be interpreted as accurate. Thus, for the *max*

strategy the overall accuracy is 0.368 for all labels and 0.530 for the top-ranked. However, *related but too broad*-labels are also meaningful in most cases, depending on the application at hand (see Section V). If we consider a combined accuracy of all *good fit*- and *too broad*-labels the value is 0.506; the combined accuracy of the labels at rank one is 0.772. The numbers for the category *related but inconvenient* are less informative. Apparently, a negligible share exists (0.045 of the top ranked labels). However, this category is particularly subjective. Therefore, one has to examine individual cases instead. The table shows the distribution of the annotators assessments. Thus, lower *unrelated*-values are better. Overall, the performance of the *max strategy* is slightly better than the *top strategy* on our test set. This result demonstrates the capability of the *max strategy* to discover small topic areas and label them correctly.

The overall distributions do not take the majority decision of the annotators into account. Therefore, we introduce two additional measures. First, *Precision@k* (*P@k*) as proposed by Hulpus et al.:

$$P@k = \frac{\#Hits \text{ with rank} \leq k}{k} \tag{4}$$

It determines how many labels of the first k labels (k = [1,5]) are a *Hit*. A *Hit* is a topic that was assessed a *good fit* (or *good fit or broader* respectively) by at least two of the three annotators. The second is an adaptation of *Coverage@k* (*C@k*) used by Hulpus et al. They measured the fraction of topics that have at least one fitting label. Hulpus et al. model topics explicitly. Hence, they were able to determine *C@k* on a per-topic-level. Instead, we model topics implicitly, i.e. we determine labels for an entire utterance at once (see Subsection III-B). Thus, we adapted *C@k* to fit our set up:

$$C@k = \frac{\#\text{u. w/ at least 1 } Hit \text{ at rank} \leq k}{\#\text{utterances}} \tag{5}$$

Our adaptation determines the fraction of utterances for which our approach produces at least one *Hit* in the top-k. Since we want to determine the precision and coverage of accurate or reasonably accurate labels, we consider *good fit* and *good fit or broader* only. Plots for both are presented in Figure 2. Again, the *max strategy* outperforms the *top strategy* almost always. As expected, increasing values for k decreases precision but increases coverage. *C@k* ranges from 0.801 up to 0.954 for the *good fit or broader*-case, which is encouraging. Even for the *good fit*-only case *C@k* exceeds 0.909 at $k \geq 4$. Of course, higher precision values would be preferable. Nevertheless, our results are comparable to Hulpus et al. However, one must consider that although precision is calculated equally and values are similar, results have to be interpreted differently. As discussed before, our approach aims at labeling short, spoken utterances. In contrast to Hulpus et al. our approach saves an explicit topic modeling step and labels multiple topics at once. Therefore, precision is semantically slightly different and comparison needs to be considered with caution. A comparison to approaches for topic labeling on

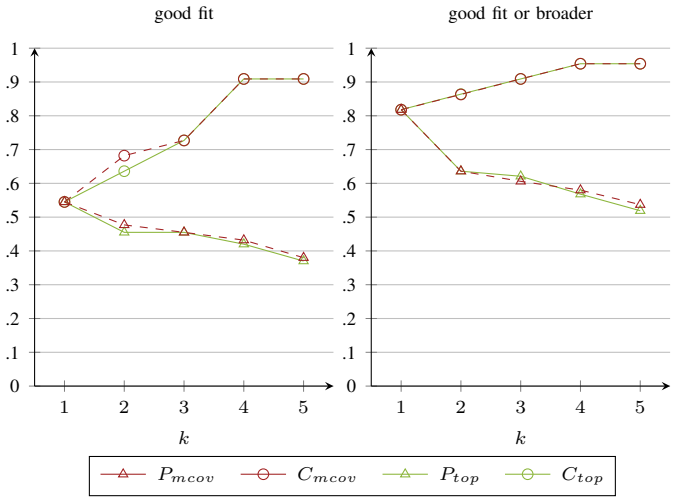

Fig. 2: *Precision@k (P)* and *Coverage@k (C)* for topics that are considered a *good fit* and "*or broader*," respectively, as achieved with the *max strategy* (mcov) and the *top strategy* (top).

spoken language can not be drawn. As discussed in Section II these approaches model topics but do not attach labels.

During evaluation, we discovered another interesting aspect: our results improve with the number of senses available. In other words, our approach has a bias towards long utterances with a broad vocabulary. This behavior is due to the graph centrality approach. It only determines meaningful labels if many connections between senses exist. In general, this is more likely the more senses contribute to the topic graph. Additionally, the homogeneity of senses has a direct influence on the performance of our two strategies. Since the *max strategy* considers all available senses, it performs better on homogeneous inputs, but is more easily diverted by discrete (irrelevant) senses. In such cases, the *top strategy* is more resilient but sometimes discards relevant senses too easily.

V. Areas of Application

In Section I we argue that unsupervised multi-topic labeling is a potential building block for a deeper understanding of spoken language. Subsequently we will justify this point by discussing areas of application for our approach.

Our work on unsupervised multi-topic labeling for spoken utterances is part of the project *PARSE* [23]. The goal of the project is to enable a layperson to program in plain spoken English. To facilitate programming with spoken language the system must understand the user's intents. Typical application areas of *PARSE* are robotics, home automation, and the like. While most of the process is independent of the domain, the target systems are modeled in ontologies (Figure 3 shows the architecture of PARSE). For the time being, *PARSE* is configured with the appropriate ontology for the use-case, e.g., a robot or a home automation system API. *PARSE* is equipped with agents for deep spoken language understand-

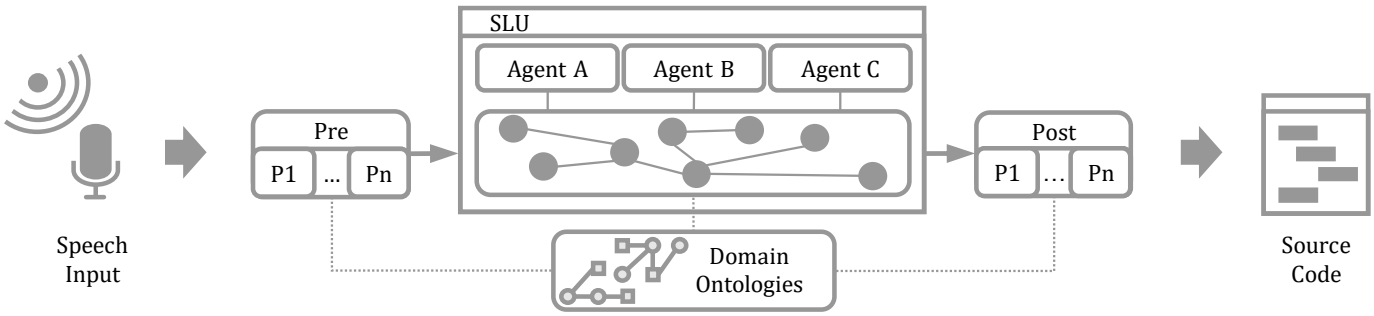

Fig. 3: The architecture of *PARSE*.

ing. SLU tasks encompass detection of actions and control structures [24], [25], analysis of coreference and context [26], or – as proposed here – topic labeling. If the graph cannot be transformed into a proper intent model, the utterance is likely to be incomplete or ambiguous. In such situations the user is queried for clarification [27].

In the upcoming subsection we discuss how we plan to improve *PARSE*'s language understanding abilities with the help of topic labeling.

A. Ontology selection

The first idea tackles the issue that albeit *PARSE* is almost domain-agnostic it still must be configured with a fitting domain ontology to work properly. First, we tried to merge all available domain ontologies. However, this method leads to ambiguities in the resulting ontology. Also the accuracies of *PARSE*'s language analyses that use the ontology diminishes. Therefore, we have to stick to small, precise ontologies.

With the help of topic labeling we might be able determine the required ontology at runtime. First, we attach topics to all ontologies. For the time being, this is a manual step. However, we plan to apply our topic labeling approach to the domain ontologies and determine topics automatically. Second, we compare the topics from the spoken utterance with the ontology topics. We select the ontologies that share the most topics with the user utterance.

We have already implemented a prototype of the ontology selection. First results are promising. Therefore, we plan to fully implement and evaluate the approach.

B. CyC (micro theories)

Another idea concerns world knowledge bases such as CyC [28]. We have experimented largely with CyC to enrich different language inputs with world knowledge, e.g. to prove the feasibility of a described course of action. However, the knowledge CyC stores is vast and hard to handle if one does not use precise queries. Luckily, all information is stored in so-called *micro theories* that cover knowledge about a certain topic. Thus, if we match topics extracted from spoken utterances with micro theories, we might be able to reduce the search space and improve querying.

C. Context modeling

In *PARSE* we build a comprehensive context model. Among other information the model includes concept relations between entities [26]. The precision of the conceptualization might improve if we incorporate information about the current topics; e.g. we might be able to distinguish the concepts *cup(dishware)* and *cup(trophy)* in more contexts.

D. Dialog interaction

PARSE also employs an extensible dialog component to resolve ambiguous situations [27]. With topic information at hand, we are able to pose more precise queries; e.g. if the system has understood that the topic is *kitchen* but missed some parts of the utterance, we might ask the user, "Do you mean 'go to the fridge'?", instead of replying, that the system has not understood the last word.

VI. Conclusion and Future Work

We have presented an approach for unsupervised multi-topic labeling that is tailored to spoken language. State-of-art approaches either depend on large textual corpora or model topics but do not attach labels. We see topic labeling as a fundamental building block to gain a deeper understanding of spoken utterances.

The contribution of the paper is two-fold. First, we have adapted the approach for word sense disambiguation by Mihalcea and Csomai [4] to short, spoken utterances. We can confirm their results; on the Wikipedia data set we achieve a similar recall (recall 0.799 vs. 0.831 in the original paper). An additional evaluation on a speech corpus with instructions for a robot shows that the method works properly on previously unseen input (F_1 0.887).

The second contribution is the approach for unsupervised multi-topic labeling for spoken utterances. Based on the word senses – i.e. the attached Wikipedia articles – we construct so-called topic graphs from DBpedia relations. We use graph centrality to determine the topics. We implemented two strategies to find the most central terms, called *top strategy* and *max strategy*. The first strategy creates topics that describe the dominant part of the utterance. The latter covers as many distinct senses as possible. Our approach creates multiple topics per utterance. Our evaluation shows that the *max strategy* slightly outperforms the *top strategy* in almost all cases.

However, the *max strategy* is more sensitive to single unrelated parts of the utterance. The overall results are promising. In a user study subjects assessed 53% of the top-ranked topic labels as *good fit*. Furthermore, for 90.9% of the evaluated utterances at least one of the top four topic labels was considered a *good fit*. If we also take labels into account that were assessed as *related but too broad* the results are even more encouraging. Subjects judged that 77.2% of the top ranked labels fit into this category; 95.4% of the utterances receive at least one topic label of this quality.

Fortunately, in most potential application areas *good or too broad*-labels are equally useful. E.g., we plan to automatically select the best fitting domain ontologies for our research project *PARSE*. Therefore, we plan to attach topics to utterances and ontologies simultaneously. Then, we can determine the ontology with most shared or related topics for the respective utterance. Consequently, *too broad*-labels are valuable, as long as ontology labels are similar or related to the utterance labels.

The same applies to the selection of CyC micro theories. Refining the conceptualization of our context model also works with broader topics.

Beyond that, we will utilize the topics for more precise dialog management and explore other areas. Furthermore, we plan to implement and evaluate additional strategies to determine central terms and experiment with differently weighted edges in sense graphs.

REFERENCES

[1] J. Boyd-Graber, Y. Hu, and D. Mimno, "Applications of topic models," *Foundations and Trends® in Information Retrieval*, vol. 11, no. 2-3, pp. 143–296, 2017.

[2] G. Tur and R. De Mori, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley & Sons Ltd, Mar. 2011.

[3] R. Mihalcea, "Using Wikipedia for Automatic Word Sense Disambiguation," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York: Association for Computational Linguistics, Apr. 2007, pp. 196–203.

[4] R. Mihalcea and A. Csomai, "Wikify!: Linking Documents to Encyclopedic Knowledge," in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, ser. CIKM '07. New York, NY, USA: ACM, 2007, pp. 233–242.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[6] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[7] S. Hingmire, S. Chougule, G. K. Palshikar, and S. Chakraborti, "Document Classification by Topic Labeling," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '13. New York, NY, USA: ACM, 2013, pp. 877–880.

[8] Q. Mei, X. Shen, and C. Zhai, "Automatic Labeling of Multinomial Topic Models," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '07. New York, NY, USA: ACM, 2007, pp. 490–499.

[9] D. Magatti, S. Calegari, D. Ciucci, and F. Stella, "Automatic Labeling of Topics," in *2009 Ninth International Conference on Intelligent Systems Design and Applications*, ser. ISDA '09. Washington, DC, USA: IEEE Computer Society, Nov. 2009, pp. 1227–1232.

[10] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsupervised graph-based topic labelling using dbpedia," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, ser. WSDM '13. New York, NY, USA: ACM, 2013, pp. 465–474. [Online]. Available: http://doi.acm.org/10.1145/2433396.2433454

[11] M. Muhr, R. Kern, and M. Granitzer, "Analysis of structural relationships for hierarchical cluster labeling," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010, pp. 178–185.

[12] K. Coursey, R. Mihalcea, and W. Moen, "Using Encyclopedic Knowledge for Automatic Topic Identification," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, ser. CoNLL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 210–218.

[13] A. Aker, E. Kurtic, A. R. Balamurali, M. Paramita, E. Barker, M. Hepple, and R. Gaizauskas, "A Graph-Based Approach to Topic Clustering for Online Comments to News," in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2016, pp. 15–29.

[14] M. Allahyari and K. Kochut, "Automatic Topic Labeling Using Ontology-Based Topic Models," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2015, pp. 259–264.

[15] A. Gourru, J. Velcin, M. Roche, C. Gravier, and P. Poncelet, "United We Stand: Using Multiple Strategies for Topic Labeling," in *Natural Language Processing and Information Systems*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2018, pp. 352–363.

[16] B. Shi, W. Lam, S. Jameel, S. Schockaert, and K. P. Lai, "Jointly Learning Word Embeddings and Latent Topics," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '17. New York, NY, USA: ACM, 2017, pp. 375–384.

[17] C. Cerisara, "Automatic discovery of topics and acoustic morphemes from speech," *Computer Speech & Language*, vol. 23, no. 2, pp. 220–239, 2009.

[18] T. J. Hazen, M. Siu, H. Gish, S. Lowe, and A. Chan, "Topic modeling for spoken documents using only phonetic information," in *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, Dec 2011, pp. 395–400.

[19] M.-h. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, "Unsupervised training of an hmm-based self-organizing unit recognizer with applications to topic classification and keyword discovery," *Computer Speech & Language*, vol. 28, no. 1, pp. 210–223, 2014.

[20] K. Coursey and R. Mihalcea, "Topic Identification Using Wikipedia Graph Centrality," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, ser. NAACL-Short '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 117–120.

[21] S. Kiesling, L. Dilley, and W. D. Raymond, "The variation in conversation (ViC) project: Creation of the Buckeye Corpus of Conversational Speech," *Ohio State University, Columbus, OH*, 2006.

[22] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Jan. 1977.

[23] S. Weigelt and W. F. Tichy, "Poster: ProNat: An Agent-Based System Design for Programming in Spoken Natural Language," in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering (ICSE)*, vol. 2, May 2015, pp. 819–820.

[24] S. Weigelt, T. Hey, and V. Steurer, "Detection of Conditionals in Spoken Utterances," in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, Jan. 2018, pp. 85–92.

[25] ——, "Detection of Control Structures in Spoken Utterances," *International Journal of Semantic Computing*, vol. 12, no. 03, pp. 335–360, Sep. 2018.

[26] S. Weigelt, T. Hey, and W. F. Tichy, "Context Model Acquisition from Spoken Utterances," in *The 29th International Conference on Software Engineering & Knowledge Engineering,*, Pittsburgh, PA, Jul. 2017, pp. 201–206.

[27] S. Weigelt, T. Hey, and M. Landhäußer, "Integrating a Dialog Component into a Framework for Spoken Language Understanding," in *Proceedings of the 6th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering*, ser. RAISE '18. Gothenburg, Sweden: ACM, 2018, pp. 1–7.

[28] D. B. Lenat, "CYC: A Large-scale Investment in Knowledge Infrastructure," *Commun. ACM*, vol. 38, no. 11, pp. 33–38, Nov. 1995.