Towards Improving Dialogue Topic Tracking Performances with Wikification of Concept Mentions

Seokhwan Kim, Rafael E. Banchs, Haizhou Li
Human Language Technology Department
Institute for Infocomm Research
Singapore 138632
{kims,rembanchs,hli}@i2r.a-star.edu.sg

Abstract

Dialogue topic tracking aims at analyzing and maintaining topic transitions in on-going dialogues. This paper proposes to utilize Wikification-based features for providing mention-level correspondences to Wikipedia concepts for dialogue topic tracking. The experimental results show that our proposed features can significantly improve the performances of the task in mixed-initiative human-human dialogues.

1 Introduction

Dialogue topic tracking aims at detecting topic transitions and predicting topic categories in on-going dialogues which address more than a single topic. Since human communications in real-world situations tend to consist of a series of multiple topics even for a single domain, tracking dialogue topics plays a key role in analyzing human-human dialogues as well as improving the naturalness of human-machine interactions by conducting multi-topic conversations.

Some researchers (Nakata et al., 2002; Lagus and Kuusisto, 2002; Adams and Martell, 2008) attempted to solve this problem with text categorization approaches for the utterances in a given turn. However, these approaches can only be effective for the cases when users mention the topic-related expressions explicitly in their utterances, because the models for text categorization assume that the proper category for each textual unit can be assigned based only on its own contents.

The other direction of dialogue topic tracking made use of external knowledge sources including domain models (Roy and Subramaniam, 2006), heuristics (Young et al., 2007), and agendas (Bohus and Rudnicky, 2003; Lee et al., 2008). While these knowledge-based methods have an advantage of dealing with system-initiative dialogues by controlling dialogue flows based on given resources, they have drawbacks in low flexibility to handle the user's responses and high costs for building the resources.

Recently, we have proposed to explore domain knowledge from Wikipedia for mixed-initiative dialogue topic tracking without significant costs for building resources (Kim et al., 2014a; Kim et al., 2014b). In these methods, a set of articles that have similar contents to a given dialogue segment are selected using vector space model. Then various types of information obtained from the articles are utilized to learn topic trackers based on kernel methods.

In this work, we focus on the following limitations of our former work in retrieving relevant concepts at a given turn with the term vector similarity between each pair of dialogue segment and Wikipedia article. Firstly, the contents of conversation could be expressed in totally different ways from the descriptions in the actual relevant articles in Wikipedia. This mismatch between spoken dialogues and written encyclopedia could bring about inaccuracy in selecting proper Wikipedia articles as sources for domain knowledge. Secondly, a set of articles that are selected by comparing with a whole dialogue segment can be limited to reflect the multiple relevances if more than one concept are actually mentioned in the segment. Lastly, lack of semantic or discourse aspects in concept retrieval could cause a limited capability of the tracker to deal with implicitly mentioned subjects.

To solve these issues, we propose to incorporate Wikification (Mihalcea and Csomai, 2007) features for building dialogue topic trackers. The goal of Wikification is resolving ambiguities and variabilities of every mention in natural language by linking the expression to its relevant Wikipedia concept. Since this task is performed using not

Proceedings of the SIGDIAL 2015 Conference, pages 124–128,
Prague, Czech Republic, 2-4 September 2015.

t	Speaker	Utterance	Topic Transition
0	Guide	How can I help you?	NONE→NONE
1	Tourist	Can you recommend some good places to visit in Singapore?	**NONE→ATTR**
	Guide	Well if you like to visit an icon of Singapore, Merlion park will be a nice place to visit.	
2	Tourist	That is a symbol for your country, right?	ATTR→ATTR
	Guide	Yes, we use that to symbolise Singapore.	
3	Tourist	Okay.	ATTR→ATTR
	Guide	The lion head symbolised the founding of the island and the fish body just symbolised the humble fishing village.	
4	Tourist	How can I get there from Orchard Road?	**ATTR→TRSP**
	Guide	You can take the red line train from Orchard and stop at Raffles Place.	
5	Tourist	Is this walking distance from the station to the destination?	TRSP→TRSP
	Guide	Yes, it'll take only ten minutes on foot.	
6	Tourist	Alright.	**TRSP→FOOD**
	Guide	Well, you can also enjoy some seafoods at the riverside near the place.	
7	Tourist	What food do you have any recommendations to try there?	FOOD→FOOD
	Guide	If you like spicy foods, you must try chilli crab which is one of our favourite dishes here.	
8	Tourist	Great! I'll try that.	FOOD→FOOD

Figure 1: Examples of dialogue topic tracking on Singapore tour guide dialogues

only surface form features, but also various types of semantic and discourse aspects obtained from both given texts and Wikipedia collection, our proposed method utilizing the results from Wikification contributes to improve the tracking performances compared to the former approaches based on dialogue segment-level correspondences.

2 Dialogue Topic Tracking

Dialogue topic tracking can be defined as a classification problem to detect where topic transitions occur and what the topic category follows after each transition. The most probable pair of topics at just before and after each turn is predicted by the following classifier:

$$f(x_t) = (y_{t-1}, y_t),$$

where x_t contains the input features obtained at a turn t, $y_t \in C$, and C is a closed set of topic categories. If a topic transition occurs at t, y_t should be different from y_{t-1}. Otherwise, both y_t and y_{t-1} have the same value.

Figure 1 shows an example of dialogue topic tracking in a given dialogue fragment on Singapore tour guide domain between a tourist and a guide. This conversation is divided into four segments, since f detects three topic transitions at t_1, t_4 and t_6. The mixed-initiative aspects are also shown in this dialogue, because the first two transitions are initiated by the tourist, while the other one is driven by the guide without any explicit requirement from the tourist. From these results, we could obtain a topic sequence of 'Attraction', 'Transportation', and 'Food'.

t	Speaker	Mention	Wikipedia Concept
1	Tourist	Singapore	Singapore
	Guide	Singapore	Singapore
		Merlion park	Merlion Park
2	Tourist	That	Merlion
		your country	Singapore
	Guide	that	Merlion
		Singapore	Singapore
4	Tourist	there	Merlion Park
		Orchard Road	Orchard Road
	Guide	red line train	North South MRT Line
		Orchard	Orchard MRT Station
		Raffles Place	Raffles Place MRT Station
5	Tourist	the station	Raffles Place MRT Station
		the destination	Merlion Park
6	Guide	seafoods	Seafood
		the riverside	Singapore River
		the place	Merlion Park
7	Tourist	there	Singapore River
	Guide	chilli crab	Chilli crab
		here	Singapore

Figure 2: Examples of Wikification on Singapore tour guide dialogues

3 Wikification of Concept Mentions in Spoken Dialogues

Wikification aims at linking mentions to the relevant entries in Wikipedia. As shown in the examples in Figure 2 for the dialogue in Figure 1, this task is performed by dealing with co-references, ambiguities, and variabilities of the mentions.

Following most previous work on Wikification (Bunescu and Pasca, 2006; Mihalcea and Csomai, 2007; Milne and Witten, 2008; Dredze et al., 2010; Han and Sun, 2011; Chen and Ji, 2011), this work also takes a supervised learning to rank algorithm for determining the most relevant concept for each mention in transcribed utterances.

In this work, every noun phrase in a given dialogue session is defined as a single mention. To capture more abstract concepts, we take not only named entities or base noun phrases, but also every complex or recursive noun phrase in a dialogue as the instance to be linked. For each mention, a set of candidates are retrieved from a Lucene [1] index on the whole Wikipedia collection divided by section-level. The ranking score $s(m, c)$ for a given pair of a mention m and its candidate concept c is assigned as follows:

$$s(m,c) = \begin{cases} 4 & \text{if } c \text{ is the exactly same as } g(m), \\ 3 & \text{if } c \text{ is the parent article of } g(m), \\ 2 & \text{if } c \text{ belongs to the same article} \\ & \text{but different section of } g(m), \\ 1 & \text{otherwise.} \end{cases},$$

where $g(m)$ is the manual annotation for the most relevant concept of m.

[1] http://lucene.apache.org/

Name	Description
SP	the speaker who spoke that mention
WM	word n-grams within the surface of m
WT	word n-grams within the title of c
EMT	whether the surface of m is same as the title of c
EMR	whether the surface of m is same as one of re-directions to c
MIT	whether the surface of m is a sub-string of the title of c
TIM	whether the title of c is a sub-string of the m's surface form
MIR	whether the surface of m is a sub-string of a re-directed title to c
RIM	whether a re-directed title to c is a sub-string of the m's surface form
PMT	similarity score based on edit distance between the surface of m and the title of c
PMR	maximum similarity score between the surface of m and the redirected titles to c
OC	whether c previously occurred in the full dialogue history
OC_w	whether c occurred within w previous turns with $w \in \{1, 3, 5, 10\}$

Table 1: List of features for training the ranking SVM model for Wikification

Then, a ranking SVM (Joachims, 2002) model, a pairwise ranking algorithm learned from the ranked lists, is trained based on the scores and the features in Table 1. In the execution time, the top-ranked item in the list of candidates scored by this model is considered as the result of Wikification for a given mention.

4 Wikification-based Features for Dialogue Topic Tracking

Following our previous work (Kim et al., 2014a; Kim et al., 2014b), the classifier f for dialogue topic tracking is trained on the labeled dataset using supervised machine learning techniques.

The simplest baseline is to learn the classifier based on the vector space model (Salton et al., 1975) considering bag-of-words for the terms within the given utterances. An instance for each turn is represented by a weighted term vector defined as follows:

$$\phi(x) = \left(\alpha_1, \alpha_2, \cdots, \alpha_{|W|}\right) \in R^{|W|},$$

where $\alpha_i = \sum_{j=0}^{h} \left(\lambda^j \cdot tfidf(w_i, u_{(t-j)})\right)$, u_t is the utterance mentioned in a turn t, $tfidf(w_i, u_t)$ is the product of term frequency of a word w_i in u_t and inverse document frequency of w_i, λ is a decay factor for giving more importance to more recent turns, $|W|$ is the size of word dictionary, and h is the number of previous turns considered as dialogue history features.

To overcome the limitations caused by lack of semantic or domain-specific aspects in the first baseline, we previosly proposed (Kim et al., 2014b) to leverage on Wikipedia as an external knowledge source with an extended feature space defined by concatenating the concept space with the previous term vector space as follows:

$$\phi'(x) = \left(\alpha_1, \alpha_2, \cdots, \alpha_{|W|}, \beta_1, \beta_2, \cdots, \beta_{|D|}\right),$$

where $\phi'(x) \in R^{|W|+|C|}$, β_i is the semantic relatedness between the input x and the concept in the i-th Wikipedia article and $|C|$ is the number of concepts in the Wikipedia collection. The value for β_i is computed with the cosine similarity between term vectors as follows:

$$\beta_i = \text{sim}(x, c_i) = \cos(\theta) = \frac{\phi(x) \cdot \phi(c_i)}{|\phi(x)||\phi(c_i)|},$$

where $\phi(c_i)$ is the term vector composed from the i-th Wikipedia concept in the collection.

In this work, the results of Wikification described in Section 3 are utilized to extend the feature space for training the topic tracker, instead of or in addition to the above mentioned feature values obtained from dialogue segment-level analyses. A value γ_i in the new feature space is defined as the weighted sum of the number of mentions linked to a given concept c_i within a dialogue segment as follows:

$$\gamma_i = \sum_{j=0}^{h} \left(\lambda^j \cdot \left|\left\{m_k \in u_{(t-j)} | g(m_k) = c_i\right\}\right|\right),$$

where m_k is the k-th mention in a given utterance u, $g(m)$ is the top-ranked result of Wikification for the mention m, λ is a decay factor, and h is the window size for considering dialogue history.

5 Evaluation

To demonstrate the effectiveness of our proposed approach for dialogue topic tracking using Wikification results, we performed experiments on the Singapore tour guide dialogues which consists of 35 sessions collected from human-human conversations between tour guides and tourists. All the recorded dialogues with the total length of 21 hours were manually transcribed, then these 31,034 utterances were manually annotated with the following nine topic categories: Opening, Closing, Itinerary, Accommodation, Attraction, Food, Transportation, Shopping, and Other.

	Schedule: All				Schedule: Tourist Turns				Schedule: Guide Turns			
	Transition			Turn	Transition			Turn	Transition			Turn
Features	P	R	F	ACC	P	R	F	ACC	P	R	F	ACC
α	42.08	53.48	47.10	67.97	41.88	52.59	46.63	67.15	41.96	52.11	46.49	67.13
α, β	42.12	53.38	47.08	67.98	41.84	52.75	46.67	67.08	41.91	52.03	46.42	67.13
α, γ	47.36	50.19	48.73	72.38	46.58	51.09	48.73	71.99	47.10	48.44	47.76	71.94
α, β, γ	47.35	50.24	48.75	72.43	46.57	51.09	48.72	71.99	47.02	48.21	47.61	71.93
α, γ'	50.77	49.36	50.06	79.12	50.51	49.58	50.04	81.10	50.94	49.10	50.00	78.92
α, β, γ'	50.82	49.41	50.10	79.15	50.43	49.58	50.00	81.10	50.98	49.02	49.98	78.92

Table 2: Comparisons of the topic tracking performances with different combinations of features

For topic tracking, an instance for both training and prediction of topic transition was created for every utterance in the dialogues. For each instance x, the term vector $\phi(x)$ was generated with the α values from utterances within the window sizes $h = 2$ for the current and previous turns and $h = 10$ for the history turns. The β values for representing the segment-level relevances were computed based on 3,155 Singapore-related articles which were used in our previous work (Kim et al., 2014b).

For Wikification, all the utterance were preprocessed by Stanford CoreNLP toolkit [2], firstly. Each noun phrase in the constituent trees provided by the parser was considered as an instance for Wikification and manually annotated with the corresponding concept in Wikipedia. For every mention, we retrieved top 100 candidates from the Lucene index based on the Wikipedia database dump as of January 2015 which has 4,797,927 articles and 25,577,464 sections in total and added one more special candidate for NIL detection. Then, a ranking function using SVM^{rank}[3] was trained on this dataset, which achieved 38.04, 31.97, and 34.74 in precision, recall, and F-measure, respectively, in the evaluation for Wikification for each mention-level based on five-fold cross validation. The γ values in our proposed approach were assigned based on the top-ranked results from this ranking fuction for the mentions in the dialogues.

In this evaluation, the following three different schedules were applied for both training the models and prediction the topic transitions: (a) taking every utterance regardless of the speaker into account; (b) considering only the turns taken by the tourists; and (c) by the guides. While the first schedule aims at learning the human behaviours in topic tracking from the third person point of view, the others could show the tracking capabilities of the models as a sub-component in the dialogue system which act as a guide and a tourist, respectively.

The SVM models were trained using SVM^{light} [4] (Joachims, 1999) with different combinations of the features. All the evaluations were done in five-fold cross validation to the manual annotations with two different metrics: one is accuracy of the predicted topic label for every turn, and the other is precision/recall/F-measure for each event of topic transition occurred either in the answer or the predicted result.

Table 2 compares the performances of the feature combinations for each schedule. While the dialogue segment-level β features failed to show significant improvement compared to the baseline only with term vectors, the models with our proposed Wikification-based features γ achieved better performances in both transition and turn-level evaluations for all the schedules. The further enhancement led by the oracle features with the manual annotations for Wikification represented by γ' indicates that the overall performances could be improved by refining the Wikification model.

6 Conclusions

This paper presented a dialogue topic tracking approach using Wikification-based features. This approach aimed to incorporate more detailed information regarding the correspondences between a given dialogue and Wikipedia concepts. Experimental results show that our proposed approach helped to improve the topic tracking performances compared to the baselines. For future work, we plan to apply the kernel methods proposed in our previous work also on the feature spaces based on Wikification as well as to improve the Wikification model itself for achieving better overall performances in dialogue topic tracking.

[2] http://nlp.stanford.edu/software/corenlp.shtml

[3] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

[4] http://svmlight.joachims.org/

References

P. H. Adams and C. H. Martell. 2008. Topic detection and extraction in chat. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing*, pages 581–588.

D. Bohus and A. Rudnicky. 2003. Ravenclaw: dialog management using hierarchical task decomposition and an expectation agenda. In *Proceedings of the European Conference on Speech, Communication and Technology*, pages 597–600.

Razvan C. Bunescu and Marius Pasca. 2006. Using Encyclopedic Knowledge for Named entity Disambiguation. In *EACL*, volume 6, pages 9–16.

Zheng Chen and Heng Ji. 2011. Collaborative ranking: A case study on entity linking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 771–781.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity Disambiguation for Knowledge Base Population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 277–285, Stroudsburg, PA, USA.

Xianpei Han and Le Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 945–954. Association for Computational Linguistics.

T. Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142.

S. Kim, R. E. Banchs, and H. Li. 2014a. A composite kernel approach for dialog topic tracking with structured domain knowledge from wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 19–23.

S. Kim, R. E. Banchs, and H. Li. 2014b. Wikipedia-based kernels for dialogue topic tracking. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 131–135.

K. Lagus and J. Kuusisto. 2002. Topic identification in natural language dialogues using neural networks. In *Proceedings of the 3rd SIGdial workshop on Discourse and dialogue*, pages 95–102.

C. Lee, S. Jung, and G. G. Lee. 2008. Robust dialog management with n-best hypotheses using dialog examples and agenda. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 630–637.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518.

T. Nakata, S. Ando, and A. Okumura. 2002. Topic detection based on dialogue history. In *Proceedings of the 19th international conference on Computational linguistics (COLING)*, pages 1–7.

S. Roy and L. V. Subramaniam. 2006. Automatic generation of domain models for call centers from noisy transcriptions. In *Proceedings of COLING/ACL*, pages 737–744.

G. Salton, A. Wong, and C.S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

S. Young, J. Schatzmann, K. Weilhammer, and H. Ye. 2007. The hidden information state approach to dialog management. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 149–152.