Anecdote Recognition and Recommendation

Wei Song[†], **Ruiji Fu**[‡], **Lizhen Liu**[†], **Hanshi Wang**[†], **Ting Liu**[§]
[†]Information Engineering, Capital Normal University, Beijing
[‡]Iflytek Research Beijing, Beijing
[§]Harbin Institute of Technology, Harbin
{wsong, lzliu, hswang}@cnu.edu.cn, rjfu@iflytek.com, tliu@ir.hit.edu.cn

Abstract

We introduce a novel task *Anecdote Recognition and Recommendation*. An anecdote is a story with a point revealing account of an individual person. Recommending proper anecdotes can be used as evidence to support argumentative writing or as a clue for further reading.

We represent an anecdote as a structured tuple — $< person, story, implication >$. Anecdote recognition runs on archived argumentative essays. We extract narratives containing events of a person as the anecdote story. More importantly, we uncover the anecdote implication, which reveals the meaning and topic of an anecdote. Our approach depends on discourse role identification. Discourse roles such as *thesis*, *main ideas* and *support* help us locate stories and their implications in essays. The experiments show that informative and interpretable anecdotes can be recognized. These anecdotes are used for anecdote recommendation. The anecdote recommender can recommend proper anecdotes in response to given topics. The anecdote implications contribute most for bridging user interested topics and relevant anecdotes.

1 Introduction

Building technical tools to assist learning and writing is of great significance and challenging. While a number of tools have been developed for giving feedback on spelling (Bangert-Drowns, 1993), grammar patterns (Yen et al., 2015) and organization (Burstein et al., 2003b), little exists to provide support during planning and composition process. During the process, an automated system, that can effectively collect topic oriented evidence and reading materials, will greatly reduce the cognitive load.

This paper introduces the ***anecdote recognition and recommendation*** task. An anecdote is a story with a point revealing account of an individual person or an incident. We aim to recognize anecdotes from texts and recommend anecdotes according to given topics. The recommended anecdotes can be used as evidence to support argumentative writing.

The argumentative essay is a genre of writing that requires the writer to investigate a topic and establish a position in a concise manner. In addition to create good claims, the quality of argument is greatly affected by the effectiveness of evidence. Finding relevant evidence is not easy, since it heavily depends on long-term accumulation of materials (from reading and observation) and the ability to retrieve and figure out right ones from memory. This process brings in great challenges for both novice and more sophisticated writers. Anecdotal evidence is one of the most commonly used evidence types (Hornikx, 2005). For a given claim, a system that can provide relevant anecdotes would help writers find good evidence and potentially improve the organization and the quality of essays. Recommending anecdotes can be the first step to recommend all types of evidence.

Recognizing anecdotes is related to previous work that extracts story-like elements from text. For example, Chambers and Jurafsky (2008) propose an unsupervised approach to extract narrative event chains from newswire text. A narrative event chain is a set of narrative events that share a common participant. However, extracting stories only is not sufficient. Suppose that we are writing an essay about *the value of life*, how can a system understand which stories are related to this topic? The stories are

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2592–2602, Osaka, Japan, December 11-17 2016.

Field	Content
Person	Steve Jobs
Story	Steve Jobs changed the digital world. He devoted his life to the digital revolution.
Implication	The value of life is to change the world.

Table 1: An extracted anecdote. The story is a narrative consisting of human-centric events and the implication is an automatically extracted text span, from which we can infer the meaning of Jobs's story.

	Topic: The value of life
Rank	Recommended Anecdotes
1	**Person**: Steve Jobs **Story**: Steve Jobs changed the digital world. Steve Jobs devoted his life to the digital revolution. **Implication**: The value of life is to change the world.
3	**Person**: Bruno **Story**: Bruno, who was burned to death, was a martyr of modern science. **Implication**: The value of life depends on its donation rather than its duration.
2	**Person**:Helen Keller **Story**:Helen Keller is known for her efforts in learning and her arduous work for the disabled. **Implication**: The life means fighting against the fate.

Table 2: An example of anecdote recommendation. Anecdotes are ranked according to the given topic.

objective facts, but the writing goals are subjective. The semantic relatedness between them are less direct as discussed in (Rinott et al., 2015). To close this gap, we have to figure out the meanings or intents that these stories want to express.

Considering the above issues, **we define an anecdote as a structured tuple, including *person*, *story* and *implication***. The *story* describes the factual information about the story of specific persons. The *implication* indicates the meanings and significance of the story. We recognize anecdotes from essays and re-use the anecdotes to assist future argumentative writing.

Our approach is based on discourse role identification, which automatically recognizes discourse roles such as the *thesis*, *main ideas*, *support* and *conclusion* in argumentative essays. These discourse roles are closely associated with the anecdote stories and implications.

To recognize anecdote stories, we propose a human-centric approach. The narratives containing events related to a shared person and playing a discourse role as *support* are extracted as an anecdote story. To recognize anecdote implications, we assume that the stance expressed by authors of essays implies the implications of anecdotes. Therefore, we choose the *thesis*, *main ideas* and *conclusion* that the stories support as their implications.

Table 3 presents an example of the extracted anecdotes. Our method can recognize in an essay that one implication of the story that *Steve Jobs changed the digital world* is that *the value of life is to change the world*, because they are identified as the discourse roles *support* and *main idea* respectively and the former supports the latter.

In this way, we recognize anecdotes from archived essays and store them to build an anecdote database. Based on this database, we can recommend anecdotes in response to user queries, which represent their interested topics. Table 2 shows an example of the results of anecdote recommendation. The suggestive anecdotes provide representative persons and brief descriptions of their stories related to the given topic. The recommendations would motivate uses to choose proper ones as evidence.

To summarize, we make the following contributions in this paper:

- We introduce the anecdote recognition and recommendation task. We explicitly define the structure of anecdotes and automatically extract factual anecdote stories and anecdote implications based on discourse role identification. The structured anecdotes are readable, interpretable and searchable. They can be recommended as potential evidence to support argumentative writing.

- Human evaluation demonstrates that accurately extracting anecdotes is indeed feasible. Moreover, the results in anecdote recommendation show that the recommended anecdotes have good relevance and usefulness. Anecdote implications contribute most for bridging anecdotes and user intent.

2 Related Work

2.1 Writing Assistance

There have been many tools providing technical support for assisting and evaluating writing at lexical and discourse levels (Bangert-Drowns, 1993; Burstein et al., 2003a; Yen et al., 2015) or based on collaboration (Noël and Robert, 2004; Nebeling et al., 2016). This paper extends existing work and proposes to recognize and recommend anecdotes for assisting argumentative writing. Our work is related to work on quote recommendation (Tan et al., 2015) and citation recommendation (He et al., 2010). But the focuses and techniques are quite different. To the best of our knowledge, our work is the first to recommend structured factual evidence to support argumentative writing.

2.2 Argumentation Mining

Argumentation mining aims to identify the components and their relationships in argumentation (Stab and Gurevych, 2014; Peldszus and Stede, 2015; Abbas and Sawamura, 2012; Lippi and Torroni, 2015; Feng and Hirst, 2011). Similar work focuses on identifying discourse roles in student essays and scientific abstracts (Burstein et al., 2003b; Guo et al., 2010). Our work focuses on recognizing anecdotes based on discourse role identification in student essays. Moreover, we are also interested in the association between roles, such as factual evidence and the arguments they support.

Our work is close to (Rinott et al., 2015), which aims to recommend evidence according to given claims. The main differences include: (1) In (Rinott et al., 2015), the system runs over dedicated manually labeled data. Instead, our approach automates all aspects of anecdote extraction and recommendation based on information extraction and discourse role identification. (2) Our approach explicitly defines and recognizes the implications of the anecdotal stories. This makes the stories more understandable and closes the semantic relatedness gap between the stories and user interested topics. (3) We focus on a specific application scenario: anecdote recommendation for argumentative writing.

2.3 Narrative Modeling

We extract anecdote stories by extracting human centric events. Story extraction is a kind of narrative modeling. A story is usually viewed as a sequence of events (Chambers and Jurafsky, 2008) based on information extraction (Etzioni et al., 2011). Much work has been done for extracting facts and events from various resources (Banko et al., 2007; Ritter et al., 2012; Bamman and Smith, 2015; Bamman et al., 2014). Similar techniques have been used in educational applications. For example, factual information in argumentative essays has be exploited for essay scoring (Klebanov and Higgins, 2012).

Our work differs from existing work in two folds: (1) We combine event extraction and discourse role identification for extracting anecdote stories. (2) We also uncover the implications of these stories in order to close the gap between objective facts and subjective human intent.

3 Data and Task

3.1 Data

The task we propose is recognizing and recommending anecdotes for assisting argumentative writing. We focus on dealing with argumentative essays, because there exist rich evidence used by the authors to support their claims. We aim to extract anecdotes from archived essays and use them as suggestions to users who are planning to write on similar topics. The users can use the recommended anecdotes as evidence directly, or view them as a clue to find more materials.

In order to recognize anecdotes that are likely to be used for supporting writing, we collect data from an online essay collection, LELE KeTang.[1] This collection contains different types of student essays such as *argumentative essays* and *narrative essays*. The types can be read through the tags of essays. We collected 16618 argumentative essays written by students from senior high school and above in Chinese.

[1] http://www.leleketang.com/zuowen/

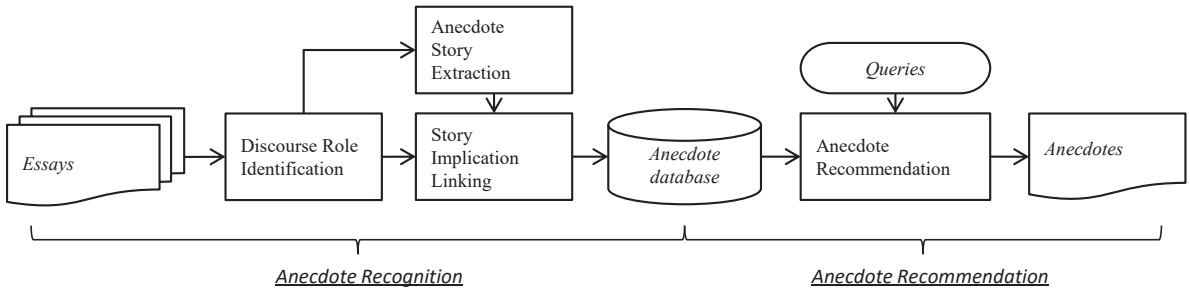

Figure 1: The framework of Anecdote Recognition and its application on Anecdote Recommendation.

3.2 Task Overview

We formally define an anecdote as a structured tuple, $< person, story, implication >$, in order to make it readable, interpretable and searchable.

An anecdote story is a concise summary of the factual events of a certain person. We view anecdote story recognition as a human-centric event extraction task. An anecdote story consists of a set of narratives describing the events of the persons.

The anecdote implication of an anecdote implies the meaning and topic of the anecdote story. We call it *implication* because the meanings or topics of the anecdote are hard to be read directly from the story itself, since the story is usually a factual description. Therefore, we have to infer the implication by means of extra information and strategies. The functions of anecdote implication should include: (1) It interprets the factual information and demonstrates how others take stance on the facts; (2) It closes the gap between objective facts and subjective writing goals to make anecdotes searchable by topics.

The general architecture of our approach is shown in Figure 1. The anecdote recognition module recognizes anecdotes from essays and stores them in a database. Anecdote stories and implications are extracted based on discourse role identification. The anecdote recommendation as an application can recommend anecdotes according to user queries.

Role	**Definition**
Introduction	introduces the background and/or grabs readers' attention
Thesis	states the main claim on the issue for which the author is arguing
Main idea	asserts ideas or aspects that are related to the thesis
Support	provides evidence to explain or support the thesis and the main ideas
Conclusion	concludes the whole essay
Other	doesn't fit into the above elements or makes no meaningful contribution

Table 3: Definitions of discourse roles.

4 Anecdote Recognition

Anecdote recognition is to extract stories and their implications from a given essay. We realize it based on discourse role identification. *Discourse roles* represent the contributions that sentences can make to text organization. Table 3 lists the discourse roles we use, which are inspired by (Burstein et al., 2003b). Our motivation is that the stories and implications relate to different discourse roles in an argumentative essay. Stories are mainly used as evidence to support the *thesis* and *main ideas* proposed by the writer. Therefore, the *thesis* and *main ideas* could be viewed as the implications of the stories in the same essay. We use the *thesis*, the *main idea* and the *conclusion* sentences as implication candidates. The reason we distinguish these roles is because they control different ranges of an essay. The *thesis* and the *conclusion* set up the main tune and conclude the whole essay, while the *main ideas* mainly control the local zones.

According to our motivation, anecdote recognition is divided into three steps: discourse role identification, story extraction and story-implication linking. Before anecdote extraction, an essay text is preprocessed by a pipeline of components including word segmentation, POS tagging, named entity

tagging, dependency parsing and semantic role labeling with HIT-LTP (Che et al., 2010). Also, we use pairs of quotations to locate quotes and view each quote as a whole.

4.1 Discourse Role Identification

We identify discourse roles based on a linear-chain Conditional Random Field (CRF) model (Lafferty et al., 2001) in order to capture the correlations among sequential predictions. The features for each sentence are mainly inspired by the previous work (Burstein et al., 2003b; Stab and Gurevych, 2014):

- **Position features** We use the relative position (beginning, middle, end) of the sentence in its paragraph and its paragraph in the essay, and the number of the sentence in the document as features.
- **Indicator features** We use manually collected cue words/phrases like *in my opinion*, *first of all* and *in conclusion* as indicators. Boolean features are designed for them.
- **Lexical features** We construct boolean features for connectives and modal verbs (such as *should*). We don't use unigrams and bigrams as features because they are sparse on a small dataset.
- **Structural features** We use the number of words, the number of clauses in the sentence and the number of sentences in the same paragraph and the ending punctuation as structural features.
- **Human and quote features** Boolean features are designed respectively to indicate whether the sentence contains human mentions, first person pronouns, third person pronouns and quotes.
- **Thesis word features** Two boolean features are used to indicate whether the sentence contains the words in essay title and the automatic extracted thesis words.

The motivation of thesis word features is that the words that reveal the thesis of an essay would help distinguish *thesis*, *main idea* and *conclusion* sentences from others. We only consider nouns, verbs and adjectives as candidate thesis words. The words in essay titles are used as thesis words. In addition, we attempt to extract thesis words automatically. We observe that the words that distribute globally in an essay tend to indicate the topic of the essay. Therefore, we rank words according to the number of paragraphs a word occurs and view the top ranked words as thesis words.

The model is learned on an annotated dataset that would be introduced in §6.1. The learned model would be used to predict the discourse roles in new essays. These identified discourse roles provide supporting information for anecdote story and implication extraction.

4.2 Story Extraction

We propose a human-centric approach to recognize anecdote stories, since anecdotes usually center around certain persons. Story extraction is conducted in two steps: recognizing human mentions, and extracting narrative events related to human mentions.

Recognizing human mentions A human mention is an observed textual reference to an individual or a group of people. Our approach considers person names, human noun phrases (NPs) and third-person pronouns. The person names are identified according to the POS tagging results provided by HIT-LTP. Human NPs refer to non-specific persons like *soldiers*. We build a dictionary containing human NPs which have a definition as HUMAN and a POS as NOUN in the common-sense knowledge base HowNet (Dong and Dong, 2006). Strings that match entries in this dictionary are marked as human NPs.

Many references to persons are in the form of pronouns. Since we have marked person names and human NPs, the pronouns should choose antecedents from them. We implement a rule-based approach considering the gender and number compatibility, syntactic roles and distance constraints. Unresolved pronouns are discarded.

Narrative event extraction Similar to (Chambers and Jurafsky, 2008), we assume that a story consists of a chain of events involving the same person. For each human mention, we extract all sentences containing it or its references. Then we conduct semantic role labeling on these sentences and extract the <agent, predicate, recipient > tuples as events. If the agent field contains the human mention, the tuple

would be retained. Since the stories have to be shown, we keep the sentences containing the retained tuples as the anecdote story candidates.

Notice that sentences that contain human mentions might be not a part of a story. Consider the following two sentences:

sentence 1 *Steve Jobs devoted his life to the digital world and designed a series of revolutionary products.*

sentence 2 *What we learn from Steve Jobs' story is that we should try our best to pursue our dreams.*

We can see that sentence 2 actually expresses an opinion rather than facts. To resolve this, we only retain the candidates with the discourse role *support* as the anecdote story.

4.3 Story Implication Linking

By now, we have extracted a set of anecdote stories. As mentioned earlier, we consider the sentences with discourse roles *thesis*, *main idea* and *conclusion* sentences as the potential implications of the stories. We have to link the stories and implications together.

We assign *thesis* and *conclusion* sentences to every anecdote story in the same essay as parts of their implications, since they cover the whole essay. The *main ideas* argument from multiple aspects. Therefore, we should link main ideas to nearby stories.

To deal with it, we train a story-idea classifier to determine whether a story and a main idea should be linked. For simplicity, we merge adjacent main idea sentences within the same paragraph as a main idea block. We derive features for every pair of a story and a main idea block. The features include (1) Paragraph distance, which is the difference between their paragraph numbers; (2) The number of shared words; (3) Connectives, since some of which like *therefore* are key indicators of the linking relation; (4) Whether they share human mentions or their references.

We train a logistic regression classifier on manually labeled story-idea pairs. For the story in a positive pair, we choose the nearest but unlinked main idea block (if there is one) to form a negative pair. During prediction, for each extracted story, we apply the classifier to determine whether it should be linked to the closest main idea blocks before and after it within 2 paragraphs. If the prediction is positive, the corresponding main idea blocks are used as part of the anecdote implication.

After story-implication linking, we get a set of <*person, story, implication*> tuples as anecdotes. By accumulating anecdotes from all available essays, we construct an anecdote database.

5 Anecdote Recommendation

Anecdote recommendation is an application of anecdote recognition. Once writers decide the main topic to address, they would attempt to find evidence to support their claims. Anecdote recommendation aims to recommend anecdotes as suggestions according to user interested topics.

The task can be described as follows: with the anecdote database C available, given a query q, the recommender returns a subset of anecdotes E, which are ranked top by a ranking model. The ranking depends on the relatedness measurement between the anecdotes and the queries.

Relatedness measurement We are interested to study which information—the anecdote story or the anecdote implication—contributes more for measuring the true relevance. Therefore, we focus on comparing the following strategies for ranking anecdotes:

- **Query-Story (QS)**. Rank based on the relatedness between the anecdote story and the query.
- **Query-Implication (QI)**. Rank based on the relatedness between the anecdote implication and the query.

To compute the semantic relatedness between the query and a text, we compare the BM25 that is a term-matching based ranking model (Robertson and Zaragoza, 2009) and a word embedding based approach (WE) — We use the average word embedding to represent a text. We are interested to see whether word embedding based approach can alleviate the semantic gap problem.

Learning to Rank Based on the above relatedness measurements, we get four ranking functions: QS-BM25, QI-BM25 and QS-WE, QI-WE. We adopt the learning to rank framework to integrate them. We

use linear ranking functions and transform the ranking problem into a two-class classification problem (Herbrich et al., 1999; Joachims, 2002). For each query, given two comparable instances whose feature vectors and labels are (x_i, y_i) and (x_j, y_j), we transform it into $(x', y') = (x_i - x_j, sign(y_i - y_j))$. After transformation, we use the SVM classifier with linear kernel to learn a ranking function.

The learned weights of some ranking functions might be negative. Since all signals should contribute positively to the final ranking, a constant is added to each weight to make sure all the weights are positive.

Training Data We build the training data based on *pseudo queries*. For an anecdote, its pseudo query is the title of the essay, from which the anecdote is extracted. It is reasonable to assume that essay titles can be used to simulate writers' search intent. We consider binary relevance labels: *relevant* and *irrelevant*. Given an essay title as a pseudo query, the anecdotes that are extracted from essays with the same title are viewed as relevant, while randomly sampled anecdotes are labeled as irrelevant to the pseudo query. In this way, we can build a training data without any manual labor.

6 Evaluation

6.1 Evaluating Discourse Role Identification

Data Two annotators manually labeled 200 student essays with discourse roles at sentence level according to a set of initial guidelines. The percentage agreement between annotators is 0.84. They discussed to reach new standards and then reviewed all annotations together. The distribution of the six discourse roles is unbalanced. The *support* discourse role accounts for 52% of all sentences, while the *introduction, thesis, main idea* and *conclusion* sentences account for 9%, 5%, 18% and 9% respectively.

Evaluation settings We conducted experiments on the corpus using 5-fold cross-validation. The precision (P), recall (R) and F_1 score are reported.

Discourse role	P	R	F_1
Introduction	73.1	74.6	73.8
Thesis	66.7	61.1	63.7
Main idea	69.0	60.9	64.6
Support	83.2	86.4	84.8
Conclusion	81.8	84.5	83.2

Table 4: Performance on identifying five discourse roles.

Results The experimental results on the corpus are shown in Table 4. We can see that our method can recognize *support* sentences well. In contrast, the *thesis* and *main ideas* are identified with moderate performance. By analyzing the errors, we found that many errors are related to the boundaries of *thesis* sentences. Some errors come from distinguishing *introduction* and *thesis*. In some cases, *introduction* sentences also involve thesis related words, although they don't explicitly make a claim, while some essays make claims at the beginning without placing any introduction. Some other errors come from incorrectly distinguishing *thesis* and *main ideas*. When there are multiple *thesis* sentences in an essay, the ones that appear later might be identified as *main idea* incorrectly. In addition, due to the imbalanced data, more sentences tend to be classified into the majority class.

6.2 The Anecdote Database

Our anecdote recognizer ran on the dataset introduced in §3 to construct an anecdote database. The database contains 26060 anecdotes, involving 9762 persons.

Data and evaluation settings We randomly sampled 200 anecdotes and asked two raters, who are students from the department of Literature, to evaluate the quality of the anecdotes. The anecdote story and implication were shown to the raters. The evaluation is based on the criteria below by judging story and implication jointly. Here is a description of the criteria:

Good	The story is understandable and complete. The implication can interpret the story.
Poor Story	The story is hard to be understood. The implication is unable to be judged.
Poor Implication	The story is understandable and complete. The implication can't interpret the story.

	Good (%)	Poor Story(%)	Poor Implication(%)
Rater1	61	22	17
Rater2	62	25	13

Table 5: The manual evaluation results by two raters.

Results Table 9 shows the evaluation results by two raters. We can see that more than 60% of the anecdotes in average are judged as *good*. This means that the anecdote stories and the corresponding implications can be effectively extracted and paired together. About 23% of the extracted anecdotes are judged as *poor story*. Most of these errors are caused by incorrectly recognized user names and the failure of pronoun resolution. 15% of the extracted anecdotes are judged as *poor implication*.

We manually labeled the story-idea pairs in 50 essays. The story-idea classifier can achieve an accuracy of 87% by cross-validation. We observe that the stories and their implications have strong locality correspondence that they tend to be within the same paragraphs.

The results show that for most recognized stories, our method can find proper implication for them. This proves that discourse role identification is a promising way for anecdote implication recognition.

6.3 Evaluating Anecdote Recommendation

6.3.1 Automated Evaluation

Data We conducted experiments on the extracted anecdote database. We stored the title of the essay from which each anecdote is extracted. There are 8141 distinct titles in all. These titles are used as queries to simulate user interested topics. We randomly sampled 1000 queries respectively to construct the training set, development set and the test set. We collected relevant instances for each query as described in §5. Each query has 3.4 relevant anecdotes in average. For each training query, we randomly sampled the same number of irrelevant instances in order to maintain a balanced training data. For each query in development set and test set, we viewed all anecdotes from the anecdotes database as candidates.

Experimental settings The parameters of SVM were tuned on the development dataset. We trained a Word2Vec model using the skip-gram algorithm with hierarchical softmax (Mikolov et al., 2013) on a dataset from Baidu Baike. The vector size is 50. The vocabulary size is 1, 825, 833. We adopt the commonly used mean average precision (MAP) and nDCG (Järvelin and Kekäläinen, 2002) as metrics.

Model	MAP	nDCG@1	nDCG@5
QS-BM25	0.357	0.406	0.387
QI-BM25	0.738	0.624	0.766
QS-WE	0.362	0.386	0.384
QI-WE	0.716	0.59	0.747
LTR	**0.744**	**0.64**	**0.786**

Table 6: The evaluation results of anecdote recommendation on pseudo queries.

Results Table 6 presents the performance of various strategies. We can see that QS-BM25 and QS-WE perform worst among all strategies. This proves that vocabulary gap exists between queries and factual story descriptions of anecdotes. In contrast, great improvements are obtained when measuring relatedness between anecdotes and queries with anecdote implications. Both QI-BM25 and QI-WE achieve much better results compared with QS-BM25 and QS-WE. This indicates that the anecdote implications play important roles in bridging user intent and anecdotes. The word embedding based approaches don't have obvious superior performance compared with conventional BM25 approaches. But by combining all signals together, LTR — that represents our learning to rank model, gains further improvement compared with any single model. This means that learning the model automatically based on pseudo query strategy is feasible.

6.3.2 Human Evaluation

Pseudo queries based evaluation indicates that proposed approach can achieve high precision on top results. To gain deeper understanding of the usefulness and interpretability, we conducted evaluation on

Topic (translation, year)
尝试(Try to do, 1994), 战胜脆弱(Overcome the weakness, 1998), 答案是丰富多彩的(The answer is rich, 2000), 心灵的选择(The Choice of heart, 2002), 转折(The turning point, 2003), 包容(Be tolerant, 2004), 我有一双隐形的翅膀(I have a pair of invisible wings, 2009), 诚信(Be honest, 2011), 深入灵魂的热爱(Deep passion, 2015).

Table 7: Essay topics for user study.

score	Description of the meanings of scores
4	The anecdote is representative and clear. It can be used as evidence directly for the given topic.
3	The anecdote is representative but not complete. It provides clues for further exploring the details.
2	The anecdote is relevant, but not representative.
1	The anecdote is irrelevant to the given topic.

Table 8: Descriptions of the meaning of each score.

manually labeled data.

Topics We chose nine argumentative topics from the recent years' college entrance examinations in Bejing, China. The topics are shown in Table 7. We only chose the topics that the essay titles are fixed so that students needn't derive their own titles according to the prompt.

Annotation For each given topics, we used the QS-BM25, QI-BM25 and LTR systems to retrieve the top 20 anecdotes respectively. We merged their results together (removing the duplicate ones) and asked 5 raters to evaluate all the results respectively using a numerical score from 1 to 4. The descriptions of the meaning of each score is shown in Table 8.

score	1	2	3	4
QS-BM25	54%	16 %	9 %	21%
QI-BM25	28%	20%	11 %	41%
LTR	26%	13%	16%	45%

Table 9: Distributions of raters' scores on manually labeled data.

6.3.3 Results

Table 9 shows the average percentage of anecdotes belonging to each score retrieved by each system. More than 70% of anecdotes extracted by LTR and QI-BM25 are relevant to the given topics. In contrast, QS-BM25 provides many more irrelevant anecdotes. This indicates that anecdote implications are better to represent the topics of the anecdotes.

Based on LTR, about 45% anecdotes are viewed as representative enough and can be used as evidence directly. A large ratio of these anecdotes are about famous people. About 16% anecdotes are considered as representative, but users have to further explore, e.g., by search, to gain more information. Parts of the low quality anecdotes are due to incorrectly recognized person names and the failure of pronoun resolution. In addition, difficult queries, like *I have a pair of invisible wings*, lead to poor performance. Such queries are kind of rhetorical device so that students have to derive the thesis themselves. Since our method is mainly based on keywords, the relevance of retrieved anecdotes is not good except that some students had used similar expressions to express their claims.

6.4 Discussions

The results show that the proposed anecdote recommender can provide relevant and representative anecdotes to user interested topics. Two main factors contribute for this. First, we use argumentative essays as data resource so that the anecdotes extracted actually have been chosen carefully by the essay authors. As a result, most of them are representative. Second, we give structures to anecdotes that usually are not considered as a structure. The anecdote implication contributes great for finding relevant anecdotes. The idea of structured retrieval can be extended to other problems as well.

On the other hand, several limitations exist. First, anecdotes extracted from student essays may be limited in narrow scopes. With the development of argumentation mining, we can extend our work to

other domains. Second, we mainly consider the relevance of anecdotes in evaluation but don't care about whether the anecdotes support or attack the given topic. Third, information beyond topical relevance such as the popularity of an anecdote or a person can be exploited. Such information is not incorporated in the current *pseudo query* based training procedure. Further relevance feedback can be conducted. Finally, organizing and diversifying the recommended anecdotes are also interesting but ignored in this study. We leave these as future work.

7 Conclusion

This paper proposes anecdote recognition and recommendation task to support argumentative writing. We not only extract concise and informative anecdote stories, but also uncover the implications of facts. The enriched structured representation makes the extracted anecdotes interpretable and searchable. Our approach is based on discourse role identification in essays. The experimental results demonstrate the effectiveness of our approach. More than 60% of our extracted anecdotes have both well described stories and interpretable implications. The extracted anecdotes can be applied for anecdote recommendation. With the help of anecdote implications, the recommender is able to suggest relevant anecdotes in response to simulated user intent. This proves that anecdote implications are useful for closing the semantic gap between factual evidence and user interested topics.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No.61402304,No.61303105), the Beijing Municipal Natural Science Foundation (No.4154065) and the Humanity & Social Science General Project of Ministry of Education (No.14YJAZH046).

References

Safia Abbas and Hajime Sawamura. 2012. Argument mining based on a structured database and its usage in an intelligent tutoring environment. *Knowledge and information systems*, 30(1):213–246.

David Bamman and Noah A Smith. 2015. Open extraction of fine-grained political statements. pages 76–85.

David Bamman, Ted Underwood, and Noah A Smith. 2014. A bayesian mixed effects model of literary character. In *ACL (1)*, pages 370–379.

Robert L Bangert-Drowns. 1993. The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational research*, 63(1):69–93.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction for the web. In *IJCAI*, volume 7, pages 2670–2676.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2003a. Criterionsm online essay evaluation: An application for automated evaluation of student essays. In *IAAI*, pages 3–10.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003b. Finding the write stuff: Automatic identification of discourse structure in student essays. *Intelligent Systems, IEEE*, 18(1):32–39.

Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL*, volume 94305, pages 789–797. Citeseer.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics.

Zhendong Dong and Qiang Dong. 2006. *HowNet and the Computation of Meaning*. World Scientific.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. 2011. Open information extraction: The second generation. In *IJCAI*, volume 11, pages 3–10.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 987–996. Association for Computational Linguistics.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 99–107. Association for Computational Linguistics.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430. ACM.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 1999. Support vector learning for ordinal regression. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, volume 1, pages 97–102. IET.

Jos Hornikx. 2005. A review of experimental research on the relative persuasiveness of anecdotal, statistical, causal, and expert evidence. *Studies in Communication Sciences*, 5(1):205–216.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

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. ACM.

Beata Beigman Klebanov and Derrick Higgins. 2012. Measuring the use of factual information in test-taker essays. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 63–72. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289. Morgan Kaufmann.

Marco Lippi and Paolo Torroni. 2015. Argument mining: A machine learning perspective. In *Theory and Applications of Formal Argumentation*, pages 163–176. Springer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Michael Nebeling, Alexandra To, Anhong Guo, Adrian A de Freitas, Jaime Teevan, Steven P Dow, and Jeffrey P Bigham. 2016. Wearwrite: Crowd-assisted writing from smartwatches. In *Proceedings of CHI*.

Sylvie Noël and Jean-Marc Robert. 2004. Empirical study on collaborative writing: What do co-authors do, use, and like? *Computer Supported Cooperative Work (CSCW)*, 13(1):63–89.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 938–948.

Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence–an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in NLP (EMNLP), Lisbon, Portugal*, pages 17–21.

Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP 2014*, pages 46–56.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2015. Learning to recommend quotes for writing. In *AAAI*, pages 2453–2459.

Tzu-Hsi Yen, Jian-Cheng Wu, Joanne Boisson, Jim Chang, and Jason Chang. 2015. Writeahead: Mining grammar patterns in corpora for assisted writing. *ACL-IJCNLP 2015*, page 139.