

Learning to Explain Ambiguous Headlines of Online News

Tianyu Liu, Wei Wei, Xiaojun Wan

Institute of Computer Science and Technology, Peking University
 The MOE Key Laboratory of Computational Linguistics, Peking University
 {ty_liu, weiwei718, wanxiaojun}@pku.edu.cn

Abstract

With the purpose of attracting clicks, online news publishers and editors use diverse strategies to make their headlines catchy, with a sacrifice of accuracy. Specifically, a considerable portion of news headlines is ambiguous. Such headlines are unclear relative to the content of the story, and largely degrade the reading experience of the audience. In this paper, we focus on dealing with the information gap caused by the ambiguous news headlines. We define a new task of explaining ambiguous headlines with short informative texts, and build a benchmark dataset for evaluation. We address the task by selecting a proper sentence from the news body to resolve the ambiguity in an ambiguous headline. Both feature engineering methods and neural network methods are explored. For feature engineering, we improve a standard SVM classifier with elaborately designed features. For neural networks, we propose an ambiguity-aware neural matching model based on a previous model. Utilizing automatic and manual evaluation metrics, we demonstrate the efficacy and the complementarity of the two methods, and the ambiguity-aware neural matching model achieves the state-of-the-art performance on this challenging task.

1 Introduction

Accuracy is one of the basic principles of journalism. However, with the prosperity of online journalism, it is increasingly hard to manage due to a lack of strict supervision and examination. A growing problem of inaccurate news headlines is especially ubiquitous. In the field of Journalism and Communication, Marquez [1980] formally points out this problem and classifies inaccurate headlines into two categories, namely ambiguous ones and misleading ones. Wei and Wan [2017] identify such headlines by utilizing machine learning methods and analyze the pervasiveness and seriousness of the problem of online clickbait, motivating us to carry out the work in this paper.

An ambiguous headline is a headline whose meaning is unclear relative to that of the content of the story [Marquez, 1980]. Ambiguous headlines make use of curiosity gap by

concealing key information of a news event. As a result, readers tend to click on the links and find out the missing information, which may degrade the reading experience of readers. For one thing, clicking the links and reading the whole passage lead to a waste of time. More seriously, the fact of the omitted information often does not live up to readers' expectations comparing with the catchy headlines. Therefore, we turn to the study of automatically explaining ambiguous news headlines by pinpointing the explanatory information in the news body.

Since the explanatory information in the news body usually focuses on the same point as the ambiguous news headline, it is intuitive to find a sentence that is most relevant to the headline. Therefore, to resolve the ambiguity¹ in an ambiguous headline, we treat the task as a variant of sentence matching problem. In recent years, a variety of sentence matching methods, including both supervised and unsupervised models, have been proposed and proved to be effective. However, in this task, general sentence matching methods select an answer without paying special attention to the ambiguous part in the headline, obtaining only limited performance.

In this paper, we define a new problem of explaining ambiguous headlines with informative texts. We build a news dataset by utilizing crawled online news and previous ambiguous headlines detection algorithm. We explore feature engineering methods as well as neural network methods. For feature engineering, we firstly build a standard SVM classifier with basic features, and then design task oriented features which take the ambiguity into special consideration to improve the performance. In order to compensate for the weakness of feature engineering methods and achieve better results, we then apply deep learning methods to our task. We propose an ambiguity-aware neural matching model based on a previous deep learning model, and it fits better with this challenging task than any other existing neural network.

Both automatic and manual evaluation metrics demonstrate the efficacy of the two methods. In addition, further experiments indicate that the feature engineering and deep learning methods are complementary to each other.

¹In this paper, 'ambiguity' or 'ambiguous part' denotes the unclear part in a headline in accordance with the definition in [Marquez, 1980].

2 Related Work

Recent years have seen some studies in the aspect of clickbait detection. Several properties and structures of clickbait headlines were investigated in [Molek-Kozakowska, 2013; 2014; Blom and Hansen, 2015], providing an entry point for preliminary automatic identification of clickbaits. Chakraborty *et al.* [2016] extracted a set of features from headlines to train a clickbait classifier. Deep learning methods are also utilized for clickbait detection. Anand *et al.* [2016] tried RNN method with word embeddings as inputs instead of hand-crafted features. Some potential non-text cues, such as user behavior analysis and image analysis, were discussed by Chen *et al.* [2015]. Biyani *et al.* [2016] further combined title-based and body-based (article informality) features to identify clickbait news. In addition to improving the performance, Wei and Wan [2017] also analyzed the pervasiveness and seriousness of the problem of clickbait. However, no previous work has investigated the problem of automatic ambiguous headline explanation.

Sentence matching is vital for many natural language tasks. Some non-neural-network methods are explored for this purpose. Bilotti *et al.* [2007] tried bag-of-word approaches with simple surface-form word matching on a sentence retrieval task and received poor results. Yih *et al.* [2013] focused on improving the lexical semantic models by performing semantic matching based on a latent word alignment structure in order to address question answering problem. Besides, different neural network models are proposed. Mikolov *et al.* [2010] presented a recurrent neural network based language model, which laid the foundation of a variety of RNN-based models. Socher *et al.* [2011] proposed an Unfolding Recursive Autoencoder for paraphrase detection. Aside from RNN, Yu *et al.* [2014] presented a bigram model based on CNN to model question and answer candidates. Yin *et al.* [2015] presented Attention Based Convolutional Neural Network which integrates mutual influence between sentences into CNN.

3 Problem Definition and Corpus

3.1 Problem Definition

Previously, Wei and Wan [2017] addressed the problem of identifying ambiguous and misleading headlines. An ambiguous headline is defined as follows [Marquez, 1980]:

An ambiguous headline is a headline whose meaning is unclear relative to that of the content of the story.

In this paper, we focus on extracting sentences from the news body to disambiguate or clarify these headlines.

It is typical of ambiguous headlines to omit some key information. The lack of knowledge arouses reader’s curiosity and lures them to click. For example:

“辽沪大战裁判主导比赛？一项刺眼数据或许才是辽宁输球真因。”

“(The referee dominated the Liaoning-Shanghai match? One shocking stat might be the real reason of the defeat of Liaoning.)”

Here, ‘一项刺眼数据’ (one shocking stat) is the exact ambiguity of the headline. The author deliberately concealed what the stat is in the headline. However, this information

can be found in the news body. We aim to extract a sentence that covers the omitted information and eliminates the confusion of the readers. In our task, the ambiguous part of the headline is pinpointed by annotators, and this information is utilized to extract more targeted sentences to explain the headlines. In practice, the readers can easily pinpoint the ambiguous parts of the headlines they encounter, and automatic detection of the ambiguous parts of headlines is not the focus of this study.

Noticing that in addition to sentence extraction, we can also try to generate a new sentence to explain the ambiguous headline, like the human annotation process, but sentence generation is not the focus of this paper, and we leave it in future work.

3.2 Corpus

We use the data set of [Wei and Wan, 2017], which contains 645 pieces of news with ambiguous headline. Additionally, we use their method to identify ambiguous headlines in the unlabeled news crawled from four major Chinese news sites (Sina, Netease, Tencent, and Toutiao) to enlarge the corpus. The final data set contains 1500 pieces of news with ambiguous headline, with 1000 for training, 100 for validation and 400 for test.

We employ six college students majoring in either Chinese or Journalism and Communication to annotate the ambiguity in the headline and write explanations accordingly. They have read relevant instructions before annotating and each piece of news is annotated by three people. The ambiguous parts labeled for each headline are mostly identical and we use the one labeled by at least two people as the annotated ambiguous part for the headline. The explanation answers written for one piece of news are seldom identical to each other, so three gold-standard explanation answers, which are allowed to be repetitive, are given for each ambiguous headline². Note that there may be multiple ambiguities in one headline, but only less than 10% of our selected news pieces have more than one ambiguous part, and for these headlines, we focus on explaining the first ambiguous part in them. Therefore, we need to produce only one explanatory sentence for each headline.

According to the definition of ambiguous news headlines, they usually deliberately omit some key information to spur curiosity. The omitted information lies in the body of the news. Therefore, in order to build training data for model learning, we use the the gold-standard sentences given by annotators as positive samples, and find negative samples by selecting from the sentences in the news body in the following way. We calculate the maximum Levenshtein ratio c between each sentence in the news body and its three corresponding gold-standard answers. We abandon those sentences whose c is greater than a threshold τ and label the rest as non-explanatory sentences (negative samples). Empirically, we set τ as 0.27. In the final training data set, we have 19711 pairs of sentences and the corresponding headlines, among which 3000 are explanatory.

²The annotators can either pick a sentence from the news body or write from scratch, in case there is not any appropriate sentence in the news.

Note that for the validation and test sets, all sentences in the news body are used for prediction, and the predicted sentence is evaluated against the human-written gold answers.

4 Feature Engineering Method

To handle this task intuitively, we first design and extract features from sentence pairs and build up a classifier based on feature engineering. Due to the definition of ambiguous news headlines, they can be explained by the news body without any additional information given. Thus, we firstly propose a set of basic features concerning only the original news headlines and body texts. However, this set of features is insensitive to the exact ambiguities of the headlines, losing sight of what a reader really wants to know. Therefore, we secondly propose ambiguity-based features, which mainly focus on the ambiguities pinpointed by human annotators. The features mentioned above are utilized to train an SVM classifier [Joachims, 2002]. We use the SVM toolkit in scikit-learn³.

4.1 Basic Features

Sentence Features

Firstly, we extract the features listed in Table 1 for each sentence in the corpus. Among these features, verbs, nouns, adjectives, and adverbs denote the motif of a sentence. The usage of idioms and slangs, named entities, numerals, sentiment, and quotations are characteristics in a sentence. We also extract the features for the headline.

We make use of “Chinese/English Vocabulary for Sentiment Analysis” released by Hownet⁴ when counting sentiment words. This vocabulary contains six files, including words expressing sentiment degree, subjective opinion, positive evaluation, negative evaluation, positive emotion and negative emotion.

Sentence Pair Features

Table 2 lists the features computed from sentence pairs. A sentence pair is composed of a sentence from the news body, and the headline associated to that article.

³scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
⁴keenage.com/

| Feature | Description |
|---------------|--|
| Nouns | Extract nouns |
| Adjectives | Extract adjectives |
| Adverbs | Extract adverbs |
| Idioms&Slangs | Extract idioms and slangs |
| Entities | Extract named entities (name, location, organization) |
| Numerals | Numerals and their headword (date, score, age) |
| Sentiment | Respectively count words expressing positive expression, negative expression, positive evaluation, and negative evaluation |
| isQuote | Assign 1 if the sentence contains: ‘:’ ‘称’ ‘回应’ ‘说’..., indicating that there is a quotation in the sentence, otherwise 0 |

Table 1: Features extracted from sentences

| Feature | Description |
|------------------|--|
| VerbSim | Compute the average cosine similarity of verbs |
| NounSim | Compute the average cosine similarity of nouns |
| AdjSim | Compute the average cosine similarity of adjectives |
| AdvSim | Compute the average cosine similarity of adverbs |
| NamedEntitySim | Count the numbers of same named entities |
| NumeralSim | Count the numbers of same numeral |
| Idioms&SlangsSim | Count the numbers of same idioms and slangs |
| SentiGap | Calculate the absolute value of sentiment difference between the two sentences. This feature is an overall difference of the positive and negative expressions mentioned in <i>Sentiment</i> |
| IsQuote | Assign 1 if both of the sentences have quotations, otherwise 0 |

Table 2: Features of sentence pairs

In a particular sentence, a set of N words (for instance, nouns) is represented as $[w_1, w_2, \dots, w_N]$, where $w_n \in \mathbb{R}^d$, corresponds to the (pretrained) d -dimensional vector presentation of the n -th word in the word set. We use the average vectors of the two sets of words to calculate their cosine similarity s .

4.2 Ambiguity-Based Features

The ambiguous part in the headline is important for the headline disambiguation task. It reflects the confusion of a reader and helps to determine an explanatory sentence that is more pertinent to the omitted information.

An ambiguous part extracted from the headline of length M is represented as $[z_1, z_2, \dots, z_M]$, where $z_n \in \mathbb{R}^d$ is the (pretrained) d -dimensional vector representation of the n -th word in the sequence. We again use the average vector \bar{z} to obtain the representation of ambiguity information:

$$\bar{z} = \frac{1}{M} \sum_{n=1}^M z_n \quad (1)$$

An author often reuses the words in the headline or words that are similar to them in the news body, in order to articulate their exact meaning. Thus, we utilize the information of ambiguity and propose two ambiguity-based features:

$$maxsim = \max_{n=1}^N s_n \quad (2)$$

$$avgsim = \frac{1}{N} \sum_{n=1}^N s_n \quad (3)$$

where N is the number of words in a news body sentence, s_n is the cosine similarity between the n -th word and \bar{z} .

Taking the ambiguity information into account, we improve the performance of the SVM classifier, which demonstrates the effectiveness of the ambiguity-based features.

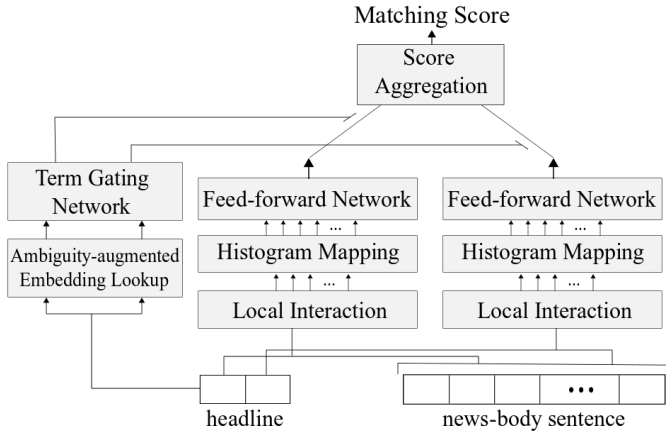


Figure 1: Architecture of the Ambiguity-aware Deep Relevance Matching Model

5 Neural Network Model

In order to fit subtler cases and achieve better results, deep learning methods are applied to and modified for our task. We carefully select and utilize a deep neural network model named Deep Relevance Matching Model (DRMM) described in [Guo *et al.*, 2016]. Figure 1 depicts the architecture of DRMM. The model first builds local interactions between each pair of terms from a sentence of the news body and its headline and obtain a matching histogram. It then employs a feed-forward matching network to learn hierarchical matching patterns and produce a matching score for each term of the headline. Finally, the overall matching score is obtained by aggregating the scores from each single term with a term gating network. The model optimizes the hinge loss with Adam.

The original paper states that DRMM focuses more on relevance rather than semantic relation between sentences, which exactly serves to our task. In the task of headline disambiguation, it is vital to find a sentence that is most relevant to the ambiguity of headline. In addition, due to the complexity and diversity of news, the semantic relations between headlines and sentences in the news body are often heterogeneous and difficult to be classified. Therefore, in our task, the semantic relation does not have enough scope for its abilities, unlike in the sentence matching tasks where the two texts are usually homogeneous such as paraphrase identification and question answering.

5.1 Matching Histogram Mapping

The input of DRMM is the local interactions between each pair of terms from a sentence of the news body and its headline, namely a matching histogram. We calculate the cosine similarity of each pair of terms and group them according to different levels of signal strengths regardless of their positions. Specifically, the local interaction is within the interval $[-1, 1]$, and we discretize the interval into a set of ordered bins and accumulate the count of local interactions in each bin. For example, if we set the bin size to 0.5, we will obtain five bins $[-1, -0.5)$, $[-0.5, 0)$, $[0, 0.5)$, $[0.5, 1)$, $[1, 1]$. Given

a headline “*football*” and a sentence in news body (*football, Brazil, foul, stadium, winter, grass*), and the corresponding local interactions are $(1, 0.2, 0.7, 0.5, -0.6, 0.1)$, we will obtain a matching histogram as $[1, 0, 2, 2, 1]$. We also apply the approach of calculating logarithm over the count value in each bin, which is proved to have the best performance in the original paper.

5.2 Term Gating Network

The DRMM model has a joint deep architecture at the headline term level, which enables itself to model the importance of the headline terms explicitly. The term gating network produces an overall relevance score by aggregating the score for each headline term. We employ the softmax function as the gating function.

$$g_n = \frac{\exp(f(x_n))}{\sum_{i=1}^N \exp(f(x_i))}, n = 1, 2, \dots, N \quad (4)$$

where x_n denotes the input of the n -th term in the headline and $f(x_n)$ denotes the output of neural network when the input is x_n . The result g_n indicates the significance of the n -th headline term.

5.3 Ambiguity-Aware DRMM (A2DRMM)

In the original paper, the model has the best performance when using the *inverse document frequency* (IDF) of headline terms as the input of the gating function. However, this method cannot capture the information of ambiguities in headlines. To make use of the information of ambiguity in a headline, we change the input to ambiguity-augmented embedding and propose the Ambiguity-aware DRMM (A2DRMM).

We concatenate the ambiguity representation \bar{z} (defined in (1)), the n -th headline term vector x_n , and its IDF h . Therefore, the ambiguity-augmented embedding of the n -th headline term with the headline ambiguity z is

$$e_n^z = \bar{z} \oplus x_n \oplus h, n = 1, 2, \dots, N \quad (5)$$

where \oplus is the vector concatenation operator. Note that the dimension of e_n^z is $(2d + 1)$. Additionally, a fully-connected network is used to learn attention signal from the input vector, which was proved to be effective for stance classification task in [Du *et al.*, 2017].

6 Experiments

In this section, we set experiments to evaluate the performance of our proposed approaches. We split the data set of 1500 pieces of news into a training set, a validation set, and a test set in the proportion of 10:1:4. As mentioned in Section 3.2, we process the training set to construct 3000 positive training samples and 16711 negative training samples. In the validation and test sets, we have a total of 2813 and 9154 sentences in news bodies, respectively, and all the sentences in a news body are candidates for the corresponding headline. In the test phase, the sentence predicted as explanatory is compared with the human-written answers with both automatic evaluation metrics and human evaluation.

6.1 Baselines

We compare our methods with two simple baselines and several neural network models which have shown good performances over sentence matching tasks. The methods are as follows:

Random: It randomly chooses a sentence from the news body. This is the lower bound method.

Similarity-based approach (SIM): It is a baseline using Bag-of-words model to calculate the cosine similarity between the headline and a sentence in the news body. We choose the sentence with the largest cosine similarity as the explanatory sentence.

ARC-I: ARC-I was proposed in [Hu *et al.*, 2014]. It is a general deep matching model that has been tested on a variety of NLP tasks including sentence completion, response matching, and paraphrase identification. It uses convolutional architecture to mine matching patterns and learn new representations for sentences.

ARC-II: ARC-II was proposed by the author of the model ARC-I. It uses convolutional architecture upon local interactions of the two sentences.

MatchPyramid (MatchPyr): It is another state-of-the-art interaction focused deep matching model proposed in [Pang *et al.*, 2016]. It uses a matching matrix whose entries represent the similarities between words of two sentences. Then, a convolutional neural network is used to learn matching patterns from the matching matrix.

6.2 Training Details

We use the Matchzoo Toolkit⁵ to implement the deep learning methods. For A2DRMM, we use 2-layer feed-forward network based on a histogram with the number of bins set to 60. For the gating network of A2DRMM, when using ambiguity-augmented embedding as input, we added a hidden layer of 5 nodes. All other parameters for training are set by default. The word embedding of all methods is the same 300-dimensional vectors trained with Word2Vec [Mikolov *et al.*, 2013].

6.3 Results

ROUGE Evaluation

We use the ROUGE-1.5.5 toolkit to perform evaluation for the results. The ROUGE metrics [Lin and Hovy, 2003] measure the quality of sentence by counting the precision, recall, and F-score of overlapping units, such as n-grams and skip grams, between a candidate sentence and gold-standard sentence. In this paper, we report the F-scores of the following metrics in the experimental results: ROUGE-1 or R-1 (unigram-based), ROUGE-2 or R-2 (bigram-based), and ROUGE-SU4 or R-SU4 (based on skip bigrams with a maximum skip distance of 4). Here, unigram refers to a single Chinese character. The comparison results of different methods are displayed in Table 3. SVM refers to the SVM classifier using only the basic features. A2SVM refers to the SVM classifier using both basic features and ambiguity-based features.

| Metric | R-1 | R-2 | R-SU4 |
|----------|----------------|----------------|----------------|
| RAND | 0.21688 | 0.10932 | 0.10834 |
| SIM | 0.30567 | 0.18511 | 0.18548 |
| ARC-I | 0.29977 | 0.18634 | 0.18487 |
| ARC-II | 0.27522 | 0.15866 | 0.15708 |
| MatchPyr | 0.32252 | 0.21738 | 0.21215 |
| SVM | 0.32040 | 0.21681 | 0.21018 |
| A2SVM | 0.32825 | 0.22452 | 0.21787 |
| A2DRMM | 0.35156 | 0.24284 | 0.23843 |

Table 3: ROUGE evaluation results of different methods

| Method | RAND | SIM | MatchPyr | A2SVM | A2DRMM |
|--------|------|------|----------|-------|-------------|
| Judge1 | 2.34 | 3.04 | 3.06 | 3.08 | 3.21 |
| Judge2 | 1.65 | 2.85 | 2.87 | 2.74 | 3.02 |

Table 4: Human evaluation results of different methods

As we can see, ARC-I and ARC-II perform badly as little semantic relations between an ambiguous headline and its explanatory sentence can be captured. The similarity-based approach also does not perform well. MatchPyramid obtains a result between SVM and A2SVM. And our A2DRMM method based on the DRMM model achieves the best results and it even performs significantly better than the strongest baselines A2SVM ($p - values < 0.05$ for pairwise T-test) and MatchPyramid ($p - values < 0.05$ for pairwise T-test).

Human Evaluation

We also perform human evaluation to further compare the efficacy of these models. 200 pieces of news were randomly chosen from the test set. For each piece of news, we compare the answers of five representative methods, namely, Random, Similarity-based approach (SIM), MatchPyramid (MatchPyr), A2SVM, and A2DRMM. Two judges majoring in Journalism and Communication were asked to rate each sentence with a numeric score from 1 to 5 according to its capability of explaining the headline. The scores are averaged across all sentences for each judge.

The results are shown in Table 4. We can see that the A2DRMM model performs significantly better than other methods ($p - values < 0.05$ for pairwise T-test). However, the comparison results between A2SVM and SIM of judge2 are not consistent with the ROUGE evaluation. We conclude that this is because when a judge evaluates the capability of a sentence to explain the headline, he concerns more about recall of information rather than precision. And the recall rate of SIM (0.39482 with the R-1 metric) is higher than A2SVM (0.35808 with the R-1 metric), which indicates that the SIM method tends to select long sentences with extra information and A2SVM tends to capture accurate sentences for headline disambiguation. By contrast, as our A2DRMM method bal-

| Metric | R-1 | R-2 | R-SU4 |
|-------------|---------|---------|---------|
| DRMM-w/o GN | 0.33508 | 0.22156 | 0.21762 |
| DRMM | 0.34880 | 0.23873 | 0.23394 |
| A2DRMM | 0.35156 | 0.24284 | 0.23843 |

Table 5: Influence of the term gating network

⁵github.com/faneshion/MatchZoo

| | |
|----------|---|
| Headline | 他们都把最好的生涯留在恒大，离开时都让恒大不赔本。 / They all gave the peak of their career to Guangzhou Evergrande, leaving Evergrande without sustaining loss after they quitted. |
| A2SVM | 埃尔克森从2012年年底时的23岁，到2016年年初离开广州恒大时的27岁，埃尔克森无疑是自己职业生涯最好的状态都留在了广州恒大。 / Elkeson was 23 when he joined Evergrande at the end of 2012, until the time he left in early 2016 at the age of 27, he definitely gave the best state of his career to Guangzhou Evergrande. |
| M-Pyr | 当时穆里奇来到广州恒大是24岁，而在2014年7月离开广州恒大时是28岁，在效力广州恒大四年左右时间里，穆里奇以“猎豹”的速度和技术帮助广州恒大拿到了3个中超冠军、1个亚冠冠军、1个超极杯冠军、1个足协杯冠军和1个中甲冠军，并荣耀中超联赛最佳射手、中超联赛最有价值球员、亚冠联赛最佳射手、亚冠联赛最有价值球员等殊荣。 / Muriqui was 24 when he joined Evergrande, 28 when he left in July of 2014. During the 4 years he played for the team, he won 3 CSL titles, 1 AFC Champion League title, 1 Super Cup, 1 CFA Cup, and 1 China League One title for his team and CSL Top Scorer and MVP, AFC Champion League Top Scorer and MVP for himself, with his speed of cheetah and football skills. |
| A2DRMM | 可以说，穆里奇、埃尔克森和金英权三人，都把自己最好的踢球年龄留给了广州恒大，同时在离开时也不会让球队赔本。 / It can be said that Muriqui, Elkeson, and Kim Young-gwon all gave their best years of football career to Guangzhou Evergrande and left without letting the team to sustain loss. |

Table 6: Samples of selected sentences by different methods for an ambiguous headline. The sentences aim to explain whom “they” refers to.

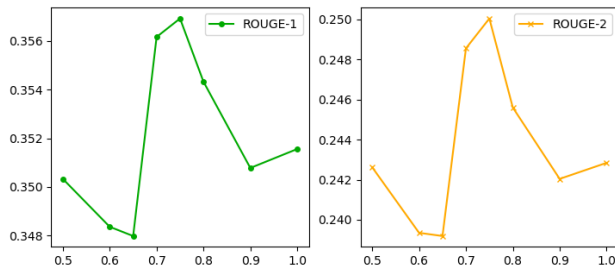


Figure 2: Results of different weights for model combination

ances the precision and recall of information, it achieves the best result in both automatic evaluation and human evaluation.

6.4 Efficacy of Term Gating Network

In order to validate the usefulness of the term gating network in DRMM, we show the comparison of three methods derived from DRMM. ‘DRMM-w/o GN’ stands for DRMM without gating network, a model whose gating network only outputs a uniform number regardless of input. We can see from the results in Table 5 that the term gating network contributes to DRMM through modeling the importance of each term in the headline, which indicates that some intrinsic importance of the terms in a headline can be discovered without any information of ambiguity. And using the ambiguity-augmented embedding further improves the results of headline explanation.

6.5 Combining A2SVM and A2DRMM

Methods based on traditional features and methods of deep learning have different but complementary properties. In this section, we seek to combine the two methods and obtain a better result. We calculate the rank of each sentence in each piece of news using SVM and A2DRMM, respectively. Then we calculate the weighted average of the two ranks, and choose the sentence with the highest rank as answer. Figure 2 shows the ROUGE results when the weight of A2DRMM is changed from 0 to 1. We can see the ROUGE scores can

be improved with proper combination weights and they reach the peak with the weight of 0.75.

6.6 Running Examples

To get a clearer view of our results, we present the sentences selected by three strong models for an ambiguous headline in Table 6. As we can see, the result of A2DRMM best explains the ambiguous part and provide enough information for readers to eliminate their confusion. The three names of persons concealed in the headline are explicitly given in the extracted sentence.

7 Conclusion

In this paper, we define a new problem of explaining ambiguous headlines with informative texts and build a benchmark dataset for evaluation. We firstly explored feature engineering methods by building a standard SVM classifier with basic features, and then designed task oriented features to improve the performance. Secondly, we explored deep learning methods and proposed an ambiguity-aware neural matching model based on DRMM. The experimental results demonstrate the efficacy of the two methods.

In future work, we will try to explore sentence generation techniques to generate new sentences to address the problem. We will further train a sequence labeling model to automatically detect the ambiguous parts in news headlines.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61772036, 61331011) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We thank the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

[Anand *et al.*, 2016] Ankesh Anand, Tanmoy Chakraborty, and Noseong Park. We used neural networks to detect clickbaits: You won’t believe what happened next! *CoRR*, abs/1612.01340, 2016.

- [Bilotti *et al.*, 2007] Matthew W Bilotti, Paul Ogilvie, Jamie Callan, and Eric Nyberg. Structured retrieval for question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 351–358. ACM, 2007.
- [Biyani *et al.*, 2016] Prakhar Biyani, Kostas Tsioutsoulouklis, and John Blackmer. ”8 amazing secrets for getting more clicks”: Detecting clickbaits in news streams using article informality. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI’16, pages 94–100. AAAI Press, 2016.
- [Blom and Hansen, 2015] Jonas Nygaard Blom and Kenneth Reinecke Hansen. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100, 2015.
- [Chakraborty *et al.*, 2016] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. *CoRR*, abs/1610.09786, 2016.
- [Chen *et al.*, 2015] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. Misleading online content: Recognizing clickbait as ”false news”. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, WMDD ’15, pages 15–19, New York, NY, USA, 2015. ACM.
- [Du *et al.*, 2017] Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. Stance classification with target-specific neural attention. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, IJCAI-17, pages 3988–3994, 2017.
- [Guo *et al.*, 2016] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, pages 55–64. ACM, 2016.
- [Hu *et al.*, 2014] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050, 2014.
- [Joachims, 2002] Thorsten Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms*, volume 186. Kluwer Academic Publishers Norwell, 2002.
- [Lin and Hovy, 2003] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics, 2003.
- [Marquez, 1980] FT Marquez. How accurate are the headlines? *Journal of Communication*, 30(3):30–36, 1980.
- [Mikolov *et al.*, 2010] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
- [Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [Molek-Kozakowska, 2013] Katarzyna Molek-Kozakowska. Towards a pragma-linguistic framework for the study of sensationalism in news headlines. *Discourse & Communication*, 7(2):173–197, 2013.
- [Molek-Kozakowska, 2014] Katarzyna Molek-Kozakowska. Coercive metaphors in news headlines a cognitive-pragmatic approach. *Brno studies in English*, 40(1), 2014.
- [Pang *et al.*, 2016] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. Text matching as image recognition. In *AAAI*, pages 2793–2799, 2016.
- [Socher *et al.*, 2011] Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in neural information processing systems*, pages 801–809, 2011.
- [Wei and Wan, 2017] Wei Wei and Xiaojun Wan. Learning to identify ambiguous and misleading news headlines. *CoRR*, abs/1705.06031, 2017.
- [Yih *et al.*, 2013] Wen-tau Yih, Ming-Wei Chang, Christopher Meek, and Andrzej Pastusiak. Question answering using enhanced lexical semantic models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1744–1753, 2013.
- [Yin *et al.*, 2015] Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *CoRR*, abs/1512.05193, 2015.
- [Yu *et al.*, 2014] Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. Deep learning for answer sentence selection. *CoRR*, abs/1412.1632, 2014.