Learning to Identify Ambiguous and Misleading News Headlines

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

Abstract

Accuracy is one of the basic principles of journalism. However, it is increasingly hard to manage due to the diversity of news media. Some editors of online news tend to use catchy headlines which trick readers into clicking. These headlines are either ambiguous or misleading, degrading the reading experience of the audience. Thus, identifying inaccurate news headlines is a task worth studying. Previous work names these headlines "clickbaits" and mainly focus on the features extracted from the headlines, which limits the performance since the consistency between headlines and news bodies is underappreciated. In this paper, we clearly redefine the problem and identify ambiguous and misleading headlines separately. We utilize class sequential rules to exploit structure information when detecting ambiguous headlines. For the identification of misleading headlines, we extract features based on the congruence between headlines and bodies. To make use of the large unlabeled data set, we apply a co-training method and gain an increase in performance. The experiment results show the effectiveness of our methods. Then we use our classifiers to detect inaccurate headlines crawled from different sources and conduct a data analysis.

1 Introduction

With the rapid development of the Internet, a variety of online news websites spring up and attract a great number of readers by providing much convenience and a wealth of information. However, the prosperity of online journalism simultaneously brings about problems including inaccuracy, unfairness, and subjectivity. We are to solve the problem of inaccurate news headlines since they are frequently complained about by readers.

With the purpose of attracting clicks, online news publishers use diverse strategies to make their headlines catchy. Generally, those clickbait headlines can be classified into two categories, namely ambiguous ones and misleading ones [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. Misleading headlines always exaggerate or distort the fact described in news bodies. Readers will discover inconsistencies after reading the whole passage, only to be left disappointed.

News with inaccurate headlines is unacceptable. It is more than a violation of professional ethics for editors. From the perspective of users, the content of this kind of news hardly lives up to their expectations comparing with the catchy headlines. Readers usually decide whether it is worth their time to read a story through headlines, so tricky headlines will waste their time and degrade the user experience. More importantly, it has been proved that headlines can shape public opinion [Tannenbaum, 1953]. According to the study, readers think differently of the accused after reading news with headlines slanted toward either innocence or guilt. From the above, we can see that it is necessary to solve the problem. Hence, we turn to the study of automatically identifying ambiguous and misleading headlines.

In this paper, we redefine the problem using Journalism and Communication knowledge. Instead of simply treating each piece of news as clickbait or non-clickbait, we separately judge whether it is ambiguous and whether it is misleading. For the ambiguous headline detection task, we mine class sequential rules (CSR) to make use of sequential information. Previous work mostly employs features extracted from headlines [Chakraborty *et al.*, 2016; Anand *et al.*, 2016]. Biyani *et al.* [2016] utilized the informality of news body, but the consistency is still underappreciated. In our method for misleading headline detection, we extract features based on headlines, bodies and the relationship between them. Since we only annotate a small part of data set, we also design a semi-supervised method, co-training, to exploit the unlabeled news. Body-independent and body-dependent features are utilized to train the sub-classifiers.

The experiment results show the effectiveness of CSR features for the ambiguous headline detection task and the consistency features for the misleading headline detection task, and performance improves when co-training is used. With the final classifiers, we classify news crawled from four major Chinese news websites, which cover different categories such as sports, society and world news. Then we conduct a data analysis upon the results.

2 Related Work

In Communication and Psychology, there have been studies on the accuracy of news headlines since decades ago. Marquez [1980] divided news headlines into three types, namely accurate, ambiguous and misleading ones, and proposed specific definitions, which are subsequently used in our classification. Ecker *et al.* [2014] analysed the effect of misinformation in news headlines and showed that such headlines lead to misconception in readers. Several properties and structures of clickbait headlines were dug out manually [Molek-Kozakowska, 2013; Molek-Kozakowska, 2014; Blom and Hansen, 2015], providing an entry point for preliminary automatic identification of clickbaits.

It was not until recent years that some studies on clickbait detection emerged in the field of artificial intelligence. Chakraborty *et al.* [2016] extracted a set of features from headlines to train a clickbait classifier. Instead of using hand-crafted features, Anand *et al.* [2016] tried RNN method with word embeddings as inputs. Some potential non-text cues, such as user behavior analysis and image analysis, were discussed but not implemented by Chen *et al.* [2015]. Biyani *et al.* [2016] further utilized both title-based and body-based (article informality) features to identify clickbait news. However, all the above works hardly considered the relationship between headlines and bodies, and focused more on the properties of headlines. In this paper, we clearly define the problem with Marquez's professional view [1980] rather than the general wording "clickbait", detecting ambiguous headlines and misleading ones separately. In the latter task, the consistency between headlines and bodies is especially taken into account.

3 Problem Definition and Corpus

3.1 Problem Definition

From the perspective of Journalism and Communication, news headlines are classified into the three categories below [Marquez, 1980]. Note that a headline can be both ambiguous and misleading. In this paper, we divide our task into two separate parts: ambiguous headline detection and misleading headline detection.

Accurate Headlines

An accurate headline is a headline that is congruent in meaning with the content of the news story.

Such headlines are non-clickbaits.

Ambiguous Headlines

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

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:

"她是曾经的世界冠军，但现在为工作发愁。"

("She once won the world championships, but now worries about making a living.")

Misleading Headlines

A misleading headline is a headline whose meaning differs from that of the content of the story.

The differences can be either subtle or obvious. Some common tactics are exaggeration, distortion, etc, most of which aim to cause a sensation. For example:

"亚洲鲤鱼已经成为美国的噩梦。"

("Asian carp has become the nightmare of America!")

In the news body of this example, the overgrowth of Asian carp due to the lack of natural predators is described. The word "nightmare" and the exclamation mark actually exaggerate the problem.

3.2 Corpus

To cover a wide range of news, we crawled a total of 40 000 articles in six different domains (domestic, world, society, entertainment, sports, and technology) from four major Chinese news sites (Sina, NetEase, Tencent, and Toutiao).

Because of the exact definition of the problem, annotators have to read through the news headlines and bodies, which is a time-consuming and demanding job. Therefore, we randomly select 2924 pieces of news and employ 6 college students majoring in Chinese to label each news headline as ambiguous or not, and also label it as misleading or not. They have read relevant instructions before annotating and each piece of news is labeled by at least 3 people. To ensure the consistency of misleading headline annotation, we abandoned 316 controversial examples (at least one annotation differs from the others). The final labeled data set contains: 645 ambiguous and 2279 non-ambiguous; 843 misleading and 1765 non-misleading.

Note that 24 000 pieces of unlabeled news are used for co-training, and a big data analysis is conducted upon the full data set. Some of the articles are not used for co-training because we continuously crawl them after the experiments.

4 Identifying Ambiguous Headlines

According to the definition of ambiguous news headlines, they usually deliberately omit some main elements of sentences to spur curiosity, which can be visually seen without reading news bodies. Thus, we firstly follow the previous practice to extract features from headlines. But the downside is that those features are mainly word-based, losing sight of sentence structures and sequential information. Thus, we secondly mine class sequential rules (CSR) and then derive CSR features. Finally, both the basic features and CSR features are utilized to train an SVM classifier [Joachims, 2002]. SVM machine learning method is selected in this task for its outperformance over other methods. We use the SVM toolkit in scikit-learn[1].

4.1 Basic Features

Table 1 lists the basic features. Those features were proved to be effective in English clickbait detection tasks [Chakraborty *et al.*, 2016]. Clickbait words are phrases or words commonly used in catchy headlines, such as "You Won't Believe" and

[1] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Feature	Description
Wordcnt	Count the number of words
Number	Count the number of numerals
Baitword	Count the number of clickbait words
Slang	Count the number of internet slang
Punctuation	Count the number of !, ? and ...
SentDegree	Respectively count 2 sets of degree adverbs expressing "很" (very) and "非常" (extremely)
SentPolar	Respectively count words expressing positive evaluation, negative evaluation, positive emotion, and negative emotion
Distance	Compute the average distance between governing and dependent words (identified by LTP parser[5])
WHword	Count the number of Chinese interrogative pronouns
ForwardRef	Count words expressing forward-reference, including demonstratives (this,that,...) and personal pronouns (he,she,it,...)

Table 1: Basic features extracted from headlines

"Will Blow Your Mind". We translate the English clickbait word list released by Downworthy[2], and manually extend the vocabulary to adapt the characteristics of Chinese. We make use of "Chinese/English Vocabulary for Sentiment Analysis" released by Hownet[3] 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. We match internet slang using the lexicons released by SogouInput[4].

4.2 Class Sequential Rules Mining

We utilize a sequential pattern mining method described in [Liu, 2007] to find language patterns of ambiguous headlines and non-ambiguous ones, and then derive features based on the sequential rules.

Let $I = \{i_1, i_2, ..., i_n\}$ be a set of items. A sequence s is an ordered list of items. s is denoted by $\langle a_1, ..., a_i, ..., a_r \rangle$, where a_i is an item in I. A sequence $s_1 = \langle a_1, a_2, ..., a_n \rangle$ is called a subsequence of $s_2 = \langle b_1, b_2, ..., b_m \rangle$ if there exist integers $1 \leq j_1 \leq j_2 \leq ... \leq j_{n-1} \leq j_n$ such that $a_1 = b_{j_1}, a_2 = b_{j_2}, ..., a_1 = b_{j_n}$.

The sequence database D is a set of pairs $\{(s_1, y_1), (s_2, y_2), ..., (s_n, y_n)\}$, where s_i is a sequence and $y_i \in Y$ is a class label. In the context of ambiguous headlines detecting, $Y = \{ambiguous, non\text{-}ambiguous\}$. A class sequential rule (CSR) is of the following form:

$$X \rightarrow y, \text{ where } X \text{ is a sequence and } y \in Y$$

An instance (s_i, y_i) **covers** the CSR if X is a subsequence of s_i. An instance (s_i, y_i) **satisfies** the CSR if X is a subsequence of s_i and $y_i = y$. The **support** (sup) of a CSR is the fraction of instances in D that satisfy the rule. The **confidence** (conf) of the rule is the proportion of instances that cover the rule which also satisfy the rule.

To explain the above definitions, we consider the following example shown in Table 2. Setting the minimum support to be 0.2 and the minimum confidence to be 0.4, one of the CSRs is $\langle 1, 4, 7 \rangle \rightarrow c_1$. The support of this rule is 0.4 since sequence 1 and 2 satisfy the rule, and the confidence is 0.67 since sequence 1, 2 and 5 cover the rule.

ID	Sequences	Class
1	$\langle 1, 4, 5, 6, 7 \rangle$	c_1
2	$\langle 1, 4, 6, 7, 9 \rangle$	c_1
3	$\langle 1, 6, 7 \rangle$	c_1
4	$\langle 2, 6, 7 \rangle$	c_2
5	$\langle 1, 3, 4, 7 \rangle$	c_2

Table 2: An example database of CSR mining

To build a sequence database, we transform the news headlines into sequences. The item set I contains 12 labels corresponding to word types mentioned in Table 1 such as WH-words and forward-reference words. I also contains 2 labels of temporal adverbs expressing "past" and "present", as well as 9 labels corresponding to different Chinese conjunctions. Temporal adverbs are used here because some headlines cause a sensation by presenting a contrast between the past and present. In Chinese, conjunctions may increase the attractiveness of a sentence by indicating antithesis, hypothesis, etc. For each word in a headline, it may be transformed into a type label. As a result, a headline, which is originally a sequence of words, is encoded into a sequence of type labels. If there is no match between a word with any items in the vocabularies, we just skip it and move on.

Take the sentence in Section 3.1 as an example. We search for each word in the lexicons, and the word "她" (she), "曾经" (once), "但" (but) and "现在" (now) are successfully matched. Consequently, the word sequence is transformed into a label sequence as follows:

$$\langle Ref, Past, But, Present \rangle$$

With selected minimum support (minsup) and minimum confidence (minconf), we utilize the CSR mining algorithm described in [Liu, 2007] to mine CSRs from the training database. Then the sequential pattern X in each CSR is treated as a feature, which is set to 1 when being contained by a headline.

After extracting features based on CSRs, we conduct experiments with CSR features to demonstrate the efficacy. We also train classifiers with all features mentioned above and see a performance improvement comparing with basic methods.

5 Identifying Misleading Headlines

A headline is considered misleading if and only if it differs from the news body. Hence, features extracted separately from headlines and bodies cannot provide enough evidence for this task. We select four groups of features which evaluate the consistency between headlines and bodies. Additionally, we utilize a bootstrapping method, co-training, to take advantage of the large set of unlabeled data.

[2]http://downworthy.snipe.net/

[3]http://www.keenage.com/

[4]http://pinyin.sogou.com/dict/

[5]http://ltp.readthedocs.io/

5.1 Features

Body-independent Features

This set of features is extracted from headlines only, similar to those included in Table 1. Note that $ForwardRef$ is abandoned here since this feature merely reflects ambiguity.

Body-dependent Features

Some of the following features are derived from news bodies, while others reflect the consistency between headlines and bodies.

1. **Informality:** We compute the frequencies of two informality indicators, namely internet slang and bait words. Additionally, the length of news bodies is also an input feature.
2. **Sentiment:** Sentiment feature consists of the frequencies of positive evaluations, negative evaluations, positive emotions, negative emotions and subjective words.
3. **InformalGap:** We calculate the absolute value of informality difference between headlines and bodies of each piece of news.
4. **SentiGap:** We calculate the absolute value of sentiment difference between headlines and bodies of each piece of news. Specifically, this feature set contains the differences in five frequencies mentioned in $Sentiment$.
5. **Similarity:** Misleading headlines often contains words that differ from those in news bodies. Thus, we firstly count the number of named entities that occur in a headline but absent in the corresponding news body. Secondly, for each word (except entities) h_i in a headline, through calculating the cosine distance of word embeddings, we find its most similar word b in body, and record this largest cosine similarity as s_i. Then the following values are used as features:

$$minSim = \min_{i=1}^{m} s_i$$

$$avgSim = \sum_{i=1}^{m} s_i/m$$

   where m is the number of words except named entities in a headline. Thirdly, tf-idf is used to compute the overall similarity between a headline and a news summary generated by PKUSUMSUM [Zhang *et al.*, 2016].
6. **Recognizing Textual Entailment(RTE):** Textual entailment is defined as follows: a text T entails another text H if the meaning of H can be inferred from the meaning of T with common background knowledge. Previous studies proposed RTE methods based on the similarity of dependency trees [Kouylekov and Magnini, 2005; Wang and Neumann, 2007]. To simplify the problem, we parse the headline and sentences in body into dependency trees, and calculate RTE-score by matching pairs of governing and dependent words which are sentence skeletons. Specifically, we search sentences in the body for dependency pairs occurred in the headline. While matching, synonym, hypernym, hyponym, and antonym are taken into account and assigned different weights. RTE-score equals the weighted sum of each match.

5.2 The Co-Training Approach

Identifying misleading headlines is a relatively complicated problem due to the variety of baiting strategies, while we only have a small set of labeled data. In order to build a robust classifier that is capable of handling different instances, we try to make full use of the larger unlabeled data set via a semi-supervised method.

Co-training [Blum and Mitchell, 1998] is a typical semi-supervised method, which demands features based on two independent views, but the independence assumption can be relaxed. Starting with the limited data set, co-training can increase the amount of labeled data by annotating the unlabeled data with two sub-classifiers. In recent years, co-training has been successfully applied to co-reference resolution [Ng and Cardie, 2003], sentiment classification [Wan, 2009], review spam identification [Li *et al.*, 2011], etc.

Since we extract body-independent and body-dependent features from each piece of news, our task exactly satisfies the essential requirement of co-training. The algorithm framework is shown in Algorithm 1. In the experiments, we balance the parameter values of p and n at each iteration to maintain the class distribution in the labeled data. Through adding confidently predicted instances, two sub-classifiers gain useful information with the help of each other. Note that the examples with conflicting labels are excluded from $N_h \cup N_b$.

Algorithm 1 Co-Training Algorithm

Given: Body-independent (headline-dependent) features F_h; body-dependent features F_b; a set of labeled news L; a set of unlabeled news U.

Loop for I iterations:

1: Learn the first classifier C_h from L based on F_h;
2: Use C_h to label news from U based on F_h;
3: Choose p positive and n negative most confidently predicted news N_h from U;
4: Learn the second classifier C_b from L based on F_b;
5: Use C_b to label news from U based on F_b;
6: Choose p positive and n negative most confidently predicted news N_b from U;
7: Remove $N_h \cup N_b$ from U;
8: Add $N_h \cup N_b$ with the corresponding labels to L.

During co-training, both the sub-classifiers will provide prediction scores for each instance. The prediction scores are normalized into $[0, 1]$. Finally, the average of the normalized values is used as the overall prediction score of each instance.

6 Evaluation Results

In this section, we set experiments to evaluate the performance of the two identification tasks. The corpus has been described in Section 3.2. The evaluation metrics are *precision*, *recall*, and *F-score*.

6.1 Ambiguous Headlines Identification

We randomly split the labeled data set into a training set and a test set in the proportion of 3:1. In the experiments, three versions of features are used for comparison:

Unigram Features: It is a baseline using the SVM classifier and the rbf kernel to identify ambiguous headlines, with unigram features provided. Unigrams here refer to Chinese words but not characters.

Basic Features: It uses the SVM classifier and the rbf kernel to identify ambiguous headlines, with only basic features provided.

CSR Features: It uses the SVM classifier and the rbf kernel with only CSR Features provided. When mining class sequential rules, we conduct experiments with different minimum support and minimum confidence. In Table 3, we list the result of $minsup$ 0.02 and $minconf$ 0.8 since this set of parameters fit well on the training set.

All Features: It uses the SVM classifier with both basic features and CSR features provided. The $minsup$ and $minconf$ for mining are the same as above.

	Precision	Recall	F-score
Unigrams	0.426	0.457	0.441
CSR Features	0.763	0.649	0.701
Basic Features	0.650	0.761	0.701
All Features	0.709	0.803	0.753

Table 3: Comparison results of methods identifying ambiguous headlines

Table 3 shows the comparison results. The method with unigram features does not perform well, because this task is different from traditional text classification tasks. The method with CSR features is proved to be effective and outperforms over basic features in precision, owing to its strictness with sentence structures. Our method with all features gets an increase in precision, recall, and F-score comparing with the method with basic features, demonstrating the contribution of CSR features. In addition, we conduct sign-test upon the predict results of basic features and all features. Sign test is a statistical method to test for consistent differences between pairs of observations. The $p\text{-}value$ of sign test is $0.0022 < 0.05$, demonstrating the significant efficacy of our method.

6.2 Misleading Headlines Identification

In this task, the labeled data set is randomly split into a training set and a test set in the proportion of 3:1. We firstly conduct supervised learning to compare the efficacy of different features on the labeled data set. For co-training, the same test set is used for evaluation, and the training set as well as the unlabeled data is used for learning.

Feature Validation

To validate the effectiveness of each feature, we conduct supervised learning with all features and then exclude each group of body-independent features to compare the performance. In addition, we list the results of the method with all body-dependent features and the method with all body-independent features. We select the SVM classifier since it performs best comparing with other methods. Rbf kernel is used in the experiments. The baseline uses unigram features and the SVM classifier. The theoretical result of random classification is also provided for comparison, which is 0.323 because of the imbalance of the data set.

	Precision	Recall	F-score
Random	0.323	0.323	0.323
Unigrams	0.428	0.456	0.443
All Features(A)	0.646	0.768	0.702
Body-dependent	0.602	0.660	0.630
Body-independent	0.637	0.716	0.674
A-Informality	0.645	0.736	0.688
A-Sentiment	0.646	0.752	0.695
A-Similarity	0.645	0.756	0.696
A-RTEscore	0.644	0.744	0.691
A-InformalGap	0.640	0.742	0.687
A-SentiGap	0.640	0.732	0.683

Table 4: Comparison results of different feature set

Table 4 shows the result of feature validation. With all features, the classifier achieves the F-score 0.702, which significantly exceed both sub-classifiers. According to the results listed, all the six groups of body-dependent features are resultful. Among all features, $InformalGap$ and $SentiGap$ are relatively important. Thus, it can be seen that the consistency between news headlines and news bodies plays an important role in misleading headline detection.

Co-Training Results

In the previous section, the efficacy of two views of features is proved. In this section, we utilize the large number of unlabeled data by leveraging the co-training method. We compare the results of co-training and the supervised method to demonstrate the suitability of co-training in our task. The sub-classifiers use the same SVM-based machine learning method mentioned in the previous section. We use different sets of parameters and the experiment results listed in Table 5 is applied to the parameter set of $p = 10, n = 20$ and iteration number $= 50$. Our proposed co-training method outperforms the supervised method with all features and gains an increase of 0.022 in F-score. We conduct sign test again and our method passes the significance test with a $p - value$ of $0.0183 < 0.05$.

	Precision	Recall	F-score
Body-dependent	0.602	0.660	0.630
Body-independent	0.637	0.716	0.674
All Features	0.646	0.768	0.702
Co-training	0.670	0.788	0.724

Table 5: Comparison results of supervised method and co-training

Parameter sensitivity: For different parameter sets, the performance of our co-training method is slightly different. We change p and n in experiments, and also observe the results in different iteration numbers. Note that n always equals $2p$ in order to keep the proportion of positive instances and negative instances. The results are shown in Figure 1. We can see that at the beginning, F-score is on the rise along

with the increasing iteration number. After 45 iterations, the results achieve relative stability with slight fluctuation. Among different growth sizes, $p = 10$ and $n = 20$ perform best on our data set. When $p = 5$ or 2, F-score increases slowly during iteration. For a larger value of $p = 20$, the performance improves obviously in 20 iterations but no longer increase then, because a large growth size is less robust to noises.

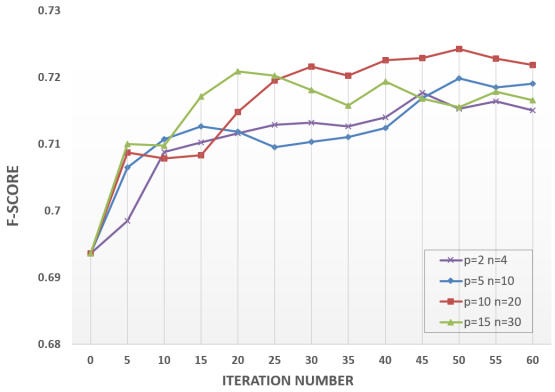

Figure 1: Influence of parameters

7 Data Analysis

We used 40 000 pieces of news crawled evenly from four major Chinese news sites (Sina, NetEase, Tencent, and Toutiao). The data covers six domains including domestic, world, society, entertainment, sports, and technology. After training the two classifiers, we identify news with ambiguous headlines and misleading ones. Then we analyze the statistics, getting some interesting and useful results.

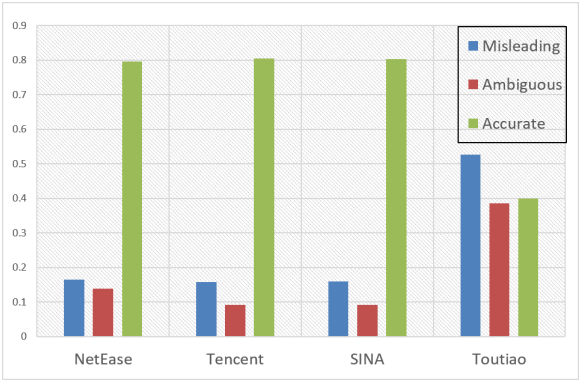

Figure 2: Comparison of different news sources

Figure 2 shows the proportion of three kinds of headlines in different news websites. Note that each three columns do not add up to 1 because there can be an intersection of the first two categories. We can see a striking contrast. Among news crawled from Sina, NetEase, and Tencent, accurate headlines make up around 80 percent, with misleading and ambiguous headlines accounting for less than 20 percent. However, only 39.9% of the news in Toutiao has accurate headlines, while misleading headlines account for more than half and ambiguous ones account for 38.6%. We can readily explain the difference. The first three websites are traditional online news organizations, while Toutiao is a kind of new media which combines various information sources without strict filtering. Lacking rigorous management, new media like Toutiao are prone to provide inaccurate information.

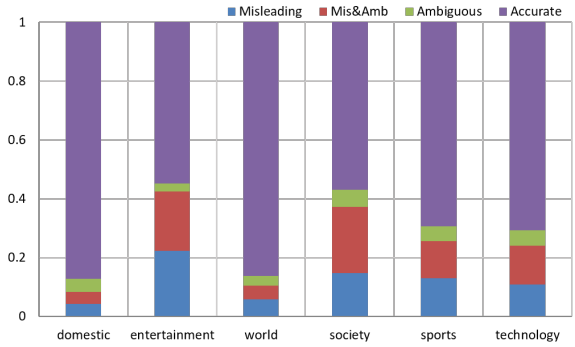

Figure 3: Comparison of different news domains

The comparison results among news in different domains are shown in Figure 3. Domestic news and world news, which usually cover political and military events, have a high rate of accuracy. The highest incidence of inaccurate headlines in entertainment news may result from its lack of seriousness. Society news telling unusual stories in social life also tends to have tricky headlines. In addition, this figure reflects that almost half the inaccurate headlines are both misleading and ambiguous.

8 Conclusion

In this paper, we study the problem of inaccurate headline detection. We crawl news online and build a Chinese data set. We divide the problem into two tasks. For ambiguous headline detection, we proposed a method based on class sequential rules and demonstrate the efficacy. For misleading headline detection, features evaluating the consistency between headlines and bodies are utilized. To exploit the larger unlabeled data set, we apply the co-training method in the latter task. The experimental results demonstrate the robustness of our identifiers.

Utilizing the final classifiers, we identify ambiguous and misleading headlines in the full data set. The results of data analysis reflect the need for more stringent regulations in journalism, especially for new media and entertainment news.

Acknowledgments

This work was supported by 863 Program of China (2015AA015403), NSFC (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 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! *arXiv preprint arXiv:1612.01340*, 2016.

[Biyani *et al.*, 2016] Prakhar Biyani, Kostas Tsioutsiouliklis, 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*, 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.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[Chakraborty *et al.*, 2016] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 9–16. IEEE, 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*, pages 15–19. ACM, 2015.

[Ecker *et al.*, 2014] Ullrich KH Ecker, Stephan Lewandowsky, Ee Pin Chang, and Rekha Pillai. The effects of subtle misinformation in news headlines. *Journal of experimental psychology: applied*, 20(4):323, 2014.

[Joachims, 2002] Thorsten Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.

[Kouylekov and Magnini, 2005] Milen Kouylekov and Bernardo Magnini. Recognizing textual entailment with tree edit distance algorithms. In *Proceedings of the First Challenge Workshop Recognising Textual Entailment*, pages 17–20, 2005.

[Li *et al.*, 2011] Fangtao Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. Learning to identify review spam. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 2488, 2011.

[Liu, 2007] Bing Liu. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.

[Marquez, 1980] FT Marquez. How accurate are the headlines? *Journal of Communication*, 30(3):30–36, 1980.

[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.

[Ng and Cardie, 2003] Vincent Ng and Claire Cardie. Weakly supervised natural language learning without redundant views. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 94–101. Association for Computational Linguistics, 2003.

[Tannenbaum, 1953] Percy H Tannenbaum. The effect of headlines on the interpretation of news stories. *Journalism Bulletin*, 30(2):189–197, 1953.

[Wan, 2009] Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*, pages 235–243. Association for Computational Linguistics, 2009.

[Wang and Neumann, 2007] Rui Wang and Günter Neumann. Recognizing textual entailment using sentence similarity based on dependency tree skeletons. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 36–41. Association for Computational Linguistics, 2007.

[Zhang *et al.*, 2016] Jianmin Zhang, Tianming Wang, and Xiaojun Wan. Pkusumsum: A java platform for multilingual document summarization. In *Proceedings of the 26th International Conference on Computational Linguistics*, 2016.