Efficient Term Frequency Inverse Document Frequency Method for Homonym Word Detection Using Concept-Based Similarity Measures

Sunil Kumar and Rajendra Gupta
Rabindranath Tagore University, Bhopal, India

Keywords: Homonym Words, TF-IDF, Concept, Weight Calculation, Concept Based Similarity Measure.

Abstract: Two or more words having the same spelling or sound but different meanings are called homonyms. The word homonyms and non-homographic homophones are complimentary subsets of homophones, which are words with the same pronunciation but different meanings. Despite their similarities, there has been substantial dispute about whether the two patterns in word recognition are similar. Identifying homonyms is one of the issues that make collecting and evaluating data from the scientific literature which is a tedious task. The terminology used to explain homonymy, heterography, and related phenomena is a bit muddled and often misunderstood, so some cleaning up is required for clarity. The paper presents a Term Frequency/Inverse Document Frequency Method for Homonym Words detection using Concept based Similarity Measures. The results show the homonym identification is achieved around 7-13 percentage better results for different datasets as compared to earlier proposed method.

1 INTRODUCTION TO HOMONYM WORDS

The word homonym is derived from the Greek word 'homonymos,' which means 'to have the same name'. The prefix 'homo' and the suffix 'nym' both indicate the same thing. As a result, homonyms are two words that have the same appearance and sound. Two or more words having the same spelling or sound but different meanings are called homonyms. These words might be perplexing at times, particularly for children learning to spell them. The word 'bat' is one of the most common examples of a homonym in English. 'Bat' is the name of an animal as well as a piece of equipment used in several sports. But when you say them out loud, they sound precisely the same, and they're spelled the same way as well (Ani et.al., 2020).

There are many homophones, or words that sound the same, in English, which can be perplexing, especially for children starting to read (Balazs et.al., 2020).

A number of words that have the same spoken form but different written forms, such as red (colour) and read (past tense). ‘Our’ and ‘hour’, ‘I’ and ‘eye’, and ‘wheat’ and ‘heat’ are some other examples. These are referred to as homophones (Bhardwaj et.al. 2018). The most significant examples are shown by (Buchta et.al. 2017) in which the homonym interpretations are employed in equal amounts.

2 METHODS FOR TEXT SIMILARITY MEASURES

The text similarity measures the compared text to existing references to determine how similar the two objects (Ferreira et.al., 2016). A number of investigations of text similarity have resulted in a variety of approaches and algorithms.

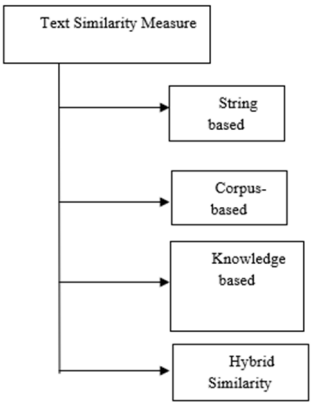

Figure 1: Techniques for Content Similarity Measures

Kumar, S. and Gupta, R.
Efficient Term Frequency Inverse Document Frequency Method for Homonym Word Detection Using Concept-Based Similarity Measures.
DOI: 10.5220/0012604300003739

In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 537-540
ISBN: 978-989-758-661-3

A semantic similarity measure and employs semantic network data to determine the degree to which words are similar. A knowledge-based similarity metric is referred to as similarity.

The most current hybrid approaches extract semantic knowledge from WordNet's structural representation as well as Internet statistic data. The author (Kim et.al., 2014) suggested TF-IDF, a new linked data metric based on a hybrid semantic similarity measure.

3 REVIEW OF EXISTING METHODS

A number of deep learning methods have been employed as a result of recent breakthroughs in the field of deep learning (Hong et.al. 2015). However, due to the lack of related metadata, such as citations and co-author information, homonym identification for a given context, such as the author's name, is confined to use in common texts, despite their qualified successes. As a result, many approaches for detecting homonyms in common texts have been devised. The author used a self-developed confusing work list to detect typographical errors and homonyms by adjusting the distance and applying a naive Bayes classifier (Hong et.al. 2015).

The aforementioned investigations, on the other hand, were conducted using a rule-based or statistical method that required an answer set, rather than relying on the semantic meaning of the word. Such methods cannot be applied to a broad text domain since the rule must be tailored to each text domain in order to get reliable results. As a result, when using the contextual word-embedding method, it is presented a novel homonym-detection technique that takes into account the semantic meaning of a word (Hong et.al. 2015). In Natural Language Processing, there are various ways for detecting word and sentence similarity (Buchta et.al. 2017).

4 EXPERIMENTAL DESIGN

The concept-based similarity metric is based on three key factors. The concepts that represent each sentence's semantic structure are the analysed tagged terms. The frequency of a concept is used to evaluate both the concept's contribution to the sentence's meaning and the main points of the document. While assessing similarity, the quantity of papers that contain the examined ideas is used to distinguish across documents. The proposed concept-based similarity measure, which considers the ctf measure to evaluate the significance of each concept at the sentence level, the tf measure at the document level, and the df measure at the corpus level, is used to evaluate these qualities.

The following aspects affect the similarity measure:

1. total number of matching ideas, called ‘m’ in the given document's verb argument structures
2. total number of sentences denoted as ‘sn’ in given document called ‘d’ which includes the matching concept denoted with ‘ci’
3. total number of labeled verb argument structures called ‘v’ in each sentence s,
4. the ctf_i of each concept ci in sentence s, where i = 1, 2,..., m for each document d
5. in each concept ci for tf_i in each document d
6. each concept's df_i
7. for each verb argument structure, the length, *Lv*, that contains a matched concept
8. in the corpus, total number of documents, N

The concept-based similarity measure between homonym words is calculated using the ctf. An exact match or a partial match between two concepts is used in concept-based matching. Both concepts share the identical homonym words, which is referred to as an exact match. A partial match occurs when one concept contains all of the words found in the other concept.

Consider the following concepts,

c_1 = ‘‘$w_1w_2w_3$’’ and c_2 = ‘‘w_1w_2’’

where c_1, c_2 are concepts and w_1, w_2, w_3 are individual words.

After removing stop words, if c_2 c_1, then c_1 holds more conceptual information than c_2. In this case, the length of c_1 is used in the similarity measure between c_1 and c_2.

The concept length is only used to compare two concepts; it has nothing to do with determining the importance of a concept in terms of sentence semantics. The ctf is used to identify relevant ideas in terms of sentence semantics known as Term Frequency (tf).

$$sim_c(d_1, d_2) = \sum_{i=1}^{m} max\left(\frac{l_{i1}}{L_{vi1}}, \frac{l_{i2}}{L_{vi2}}\right) \times weight_{i1} \times weight_{i2}, \dots\dots\dots$$

The concept-based similarity between two documents, d_1 and d_2 is calculated by:

$$weight_i = (tf\ weight_i + ctf\ weight_i) \times log\left(\frac{N}{df_i}\right)$$

In the document d, the concept-based weight of concept i, is calculated using the above equation. The terms in the above equations can be defined as :

$tf_{weight\ i}$ value represents the weight of concept i in document d.

$ctf_{weight\ i}$ value represents the weight of the defined concept i in document d at the sentence level, based on the contribution of defined concept i to the semantics of the sentences in d.

the value of $\log\left(\frac{N}{dfi}\right)$ rewards the weight of the given concept i on the corpus level, when concept i appears in a lesser number of documents.

The amount between the two values of $tf_{weight\ i}$ and $ctf_{weight\ i}$ the equation 3 represents an perfect measure of the involvement of each concept to the meaning of the sentences and to the themes mentioned in a document.

The multiplication between value of $\log\left(\frac{N}{dfi}\right)$ and value of $(tf\ weight_i + ctf\ weight_i)$ demonstrate the concepts that can be competently discriminate between documents of the entire corpus.

'a' is an advanced score, as the matching concept length methods the length of its verb argument structure, because this concept inclines to hold more conceptual data related to the meaning of its sentence.

the tf_{ij} value is normalized by the length of the text vector of the term frequency tf_{ij} in document d, where j = 1, 2, . . . , c_n and

$$tf\ weight_i = \frac{tf_{ij}}{\sqrt{\sum_{j=1}^{cn} (tf_{ij})^2}}$$

c_n is the entire number of the concepts which has a value in term-frequency in document d.

the value of ctf_{ij} is calculated by the length of the document vector of the conceptual term frequency ctf_{ij} in document d, where j = 1, 2, . . . , c_n and

$$ctf\ weight_i = \frac{ctf_{ij}}{\sqrt{\sum_{j=1}^{cn} (ctf_{ij})^2}}$$

where c_n is the total number of concepts which has a conceptual term frequency value in document *d*.

A well-known Term Frequency/Inverse Document Frequency (TF-IDF) term weighting is used with the cosine correlation similarity metric for the single-term similarity measure. The cosine measure was picked since it is frequently used in document clustering literature. It is noted that the cosine measure calculates the angle that the two document vectors make. As a result, the SIMS (single-term similarity measure) is:

$$sim_s(d_1, d_2) = cos\ cos\ (x, y) = \frac{d_1 \cdot d_2}{\|d_1\|\|d_2\|}$$

The vectors d_1 and d_2 are represented as single-term weights calculated by using the TF-IDF weighting scheme.

5 RESULTS ANALYSIS

The proposed concept-based similarity measure, which considers the ctf measure to evaluate the significance of each concept at the sentence level, the tf measure at the document level, and the df measure at the corpus level, is used to evaluate these qualities.

We have used 20news-18828 data, which is a type of real business text data to train the contextual word-embedding architecture. The 20news-18828 data came from the UCI Machine Learning Repository's News Media (NM) dataset. In addition, the Reuters-news and 20Newsgroup datasets, two public text datasets were used as 20news-18828 data in the document categorization trials. The third dataset known as Classic Dataset has been taken from UCI Machine Learning repository to check the homonyms.

The following Table 1, Table 2 and Table 3 shows the comparison between SIMS and TF-IDF methods for homonym word detection:

Table 1: Average Homonym Words Detection using SIMS and TF-IDF methods in 20news-18828 Dataset.

S.No.	Classes	No. of Words	Avg. % homonym detection	
			SIMS	TF-IDF
1	Media	750	80.40	92.20
2	Insert Error	750	78.25	85.50
3	Fixture	755	74.46	80.84
4	Safety & back-up	700	68.78	71.62
5	Network Connection	800	68.25	75.85
6	Battery and Backup	800	65.20	75..40
7	Call and Mails	800	64.43	68.40
8	User Edge	800	58.44	60.51

Table 2: Average Homonym Words Detection using SIMS and TF-IDF methods in Reuter-News Dataset.

S.No.	Classes	No. of Words	Avg. % of homonym detection	
			SIMS	TF-IDF
1	Creation	2000	82.10	96.32
2	Politics	2000	84.50	92.50
3	Technology	2000	75.41	86.80
4	Shop	2000	72.80	85.41
5	Entertainment	2000	68.64	82.70
6	Game	2000	68.85	80.80
7	Health	2000	64.92	80.35
8	Occupational	2000	63.44	80.35

When it comes in calculating concept-based relations between documents, the ctf weighting scheme is shown to be more accurate than the tf weighting scheme. The tf, ctf, and df weighting algorithms combined can accurately estimate the relevance of a concept at the sentence, document, and corpus levels, resulting in much improved clustering quality.

The Table 1 and 2 demonstrate the comparative performance between Classes of Keywords Verses No. of Data depicting Average Homonym Detection in 20news-18828, Reuter-News and Classic Dataset using SIMS and TF-IDF methods. The obtained result in Table 1 clearing depicting that the homonym identification is achieved around 6.8 percentage by applying TF-IDF methods and Table 2 shows the homonym identification is achieved around 13 percentage better results as compared to SIMS approach while in Table 3, the homonym identification is achieved around 7.78 percentage better results.

6 CONCLUSIONS

By combining the factors affecting the weights of concepts at the sentence, document, and corpus levels, a concept-based similarity measure that can accurately compute the comparison of matched texts is constructed. This makes idea matching extremely reliable and accurate, as well as concept-based similarity calculations between papers. The text clustering performance of this model significantly outperforms that of traditional single-term methods.

REFERENCES

Ani Y., Liu S. and Wang H. (2020). Error Detection in a large-scale lexical taxonomy, Information 11(2):97

Balazs J.A., Velásquez J.D. (2020). Opinion mining and information fusion: a survey, Information Fusion 27:95–110.

Bhardwaj P., Khosla P. (2018). Review of text mining techniques", IITM Journal of Management and IT, 8(1):27–31

Buchta C., Kober M., Feinerer I., Hornik K. (2017). Spherical K-Means clustering, Journal of Statistical Software, 50(10):1–22

Correia R.A., Jepson P., Malhado A.C. and Ladle RJ. (2017). Internet scientific name frequency as an indicator of cultural salience of biodiversity", Ecology Indic 78:549–555

Ferreira A.A., Veloso A., Gonçalves M.A. and Laender A.H. (2016). Self-Training author name disambiguation for information scarce scenarios", Journal of Association of Information Technology, 65(6):1257–1278

Heoy,Kang S., Seo J. (2016). Hybrid sense classification method for large-scale word sense disambiguation, IEEE Access 8:27247–27256

Hong C., Yu J., Tao D. and Wangm (2015) Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval, IEEE Transaction Ind Electron 62(6):3742–3751

Hong C., Yu J., Wan J., Tao D., Wang M. (2015). Multimodal deep autoencoder for human pose recovery, IEEE Trans Image Process 24(12):5659–5670, 2015

Kim H.K., Kim H. and Cho S. (2014). Bag-Of-Concepts: Comprehending document representation through clustering words in distributed representation, Neurocomputing, 266:336–352.