Conceptualisation of Knowledge Discovery from Web Search

Boshra F. Zopon AL_Bayaty, Dr. Shashank Joshi

Abstract— Knowledge discovery from web search is an approach to extracts knowledge from Internet using natural language processing (NLP) and search engine. Because of the inaccuracy results of keyword search in the internet, all studies of web mining method are trying to improve the accuracy or value of the information gotten from the web pages. Combination methods in WSD can do by different ways such as voting, stacking and so on. In this paper we present a Master- slave voting technique, that we will be use in our theses. This system composed of two parts as described below. We try in this way to achieve improvement by combine some slave methods in voting, to obtain high accuracy.

Index Terms— Master – Slave's technique, WSD, NLP, supervised approaches, voting, WordNet, semcore.

1 Introduction

The title of this research "**Conceptualisation of Knowledge Discovery from Web Search".** This topic is historical is one of the most topics that concern with the discover hidden and valuable knowledge using a combination of statical analysis, machine learning, search engine, and natural language processing. Evaluation any web search engine is the key to ensure the effectiveness, efficiency, Scalability, and usability of these browsing methods. DM KD KDD are broad area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence, and database systems, for the analysis of large volumes of data. There have been a large number of data mining algorithms rooted in these fields to perform different data analysis tasks. **Word sense disambiguation (WSD)**: is the task of examining word tokens in context and determining which sense of each word is being used. WSD is a task has long and rich history in computational linguistics and **Natural language processing (NLP)**, and there is robust approach to achieve high accuracy. Most of these approaches rely on contextual similarity to help choose the proper sense in a computational manner. Since a many of research for WSD, and there are a lot of supervised techniques for sense disambiguation today, a combination of such techniques could result in a highly efficient and improve the accuracy of the WSD process. In this research, we hope we can add some enhancements in search engine and natural language system fields by using combination technology of word sense disambiguation methods, we try to apply an ensemble methods (The combination strategies are called ensemble methods), which consists of main parts in figure (1), and we called this technique master- slave voting technique.

2 Objectiveof Technique

The aim of this technique is to investigate improvement in web search engine by using different supervised word sense disambiguation (WSD) classifiers; and compare supervised approaches to increase accuracy using master-slave voting technique.

3 Explain Master- Slave Voting Technique for word sense disambiguation (WSD)

Knowledge is a core or the essence component of WSD. Knowledge sources provide data which are associate senses with words. In this research we will need a WordNet or Semcor as requirements to select sense words. There are many methods were used in WSD. Some of these WSDs are well-knowing combination classifiers (methods) such as voting case, which in this case, several methods run independently and the final result is selected by voting among these methods outputs (combine several models by voting) which we called (slave classifiers, C1…to …Cn) and using it as input in next step of system[4]. Voting can be non-weighted or weighted. Weighted voting is done by adding more weight to the votes of method with higher accuracy. The biggest problem in voting is when the used techniques are similar in methodology, i.e., they make similar errors in similar situation [2] .In ***supervised WSD*** approaches use machine-learning techniques to learn a Classifier from labelled training sets, and the classifier almost called (word expert). In training and evaluation system can create data sets with proper sense for each instance. Figure (1) shown below illustrate the Master- Slave voting technique which will implements most our experiments on it, as in next subsection [2].

3.1 The experiment voting model using Master - Slave voting technique.

That will include for example a Decision list and Adaboost they will Slave methods which feed their outputs in parallel and Naïve Bayes, will be Master method, and the factor can be weighted and depending on the accuracy of slave classifiers that will tests before apply the Master- Slave technique. The Master classifier (method) will control for choice better classifier among classifier suggested by the slave lassifiers [4].

4. System Requirements

In this system the knowledge source will be WordNet or SemCor as in next subsections:

4.1 WordNet.

[Miller et al. 1990; Fellbaum 1998] which is a lexical of database. The recent version, WordNet 3.0 composed about 155,000 words and organized as synsets (more than 117,000 synsets). The synset term refers to short of synonymy set (set of words).

4.2 SemCor. SemCor

[Miller et al. 1993] which is composed words have been manually annotated with word sense from theWordNet inventory, and the original SemCor is was annotated according to WordNet 1.5. The latest versions are (e.g., 2.0, 2.1, 3.0 etc). [13].

5. Conclusion

We have described the voting system which called (**Master- Slave voting technique);** we hope that model can make enhancements in search engine field by using combination technology of knowledge-based system and web-mining methods.

6. Acknowledgment

This paper reported would not have been possible without the generous efforts of my research guide Dr. Shashank Joshi (Professor at Bharati Vidyapeeth University, College of Engineering), and support for my choices of different studies in this field.

7. Authors and Affiliations

First Auther :

Boshra F. Zopon AL_Bayaty received her B.E degree in computer science from AL_Mustansiriyah University, College of Education in 2002. And received her M.S.C degree in computer science from Iraqi Commission for Computers & Informatics, Informatics Institute for Postgraduate Studies. Doing her the PH.D. Computer Science at Bharati Vidyapeeth Deemed University, Pune. She is currently working in the Ministry of Higher Education & Scientific Research, AL_Mustansiriyah University in Iraq/ Baghdad. Her research interests include Softwere engineering and web technologies.

Second Auteur :

Shashank Joshi receive his B.E degree in Electronic and Telecommunication from Govt College of Engineering, Pune in 1988, the M.E and Ph.D. Degree in Computer Engineering from Bharati Vidyapeeth Deemed University Pune. He is currently working as the Professor in Computer Engineering Department , Bharati Vidyapeeth Deemed University, College of Engineering, Pune. His research interests include software development methodologies. He is innovatve teacher devoted to Education and Learning for the last 23 yrs.

8. References

[1] Daniel Jurafsky & James H. Martin *"Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition."* 2007.

[2] Henrich V., Reuter T. and Loftsson H. *"CombiTagger: A System for Developing Combined Taggers"*. Proceedings of the Twenty-Second International FLAIRS Conference, Sanibel Island, Florida, USA, (2009).

[3] Nitin Indurkhya and Fred J. Damerau *"HANDBOOK OF NATURAL LANGUAGE PROCESSING"* SECOND EDITION. Chapman & Hall/CRC, USA, 2010.

[4] Ahmed H. Aliwy. *"Arabic Morphosyntactic Raw Text part of Speech Tagging System"*. PhD dissertation, University of Warsaw, 2013.

[5] Facca, F. M. and Lanzi P. L., 2003, *" Recent Developments in Web Usage Mining Research*", In Kambayashi Y. et al., *Data Warehousing and Knowledge Discovery"*: 5th International Conference, DaWak 2003 Prague, September 3-5, 2003 Proceedings, Czech Republic, page 140-150.

[6] Chakrabarti, S., 2003," *Mining the web: discovering knowledge from hypertext data"*, Morgan Kaufmann Publishers, San Francisco.

[7] Hopgood, A. A., 2001, *Intelligent Systems for Engineers and Scientists*, CRC Press, Boca Raton.

[8] Miller, G. et al., *Introduction to WordNet: An On-line Lexical Database*, ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf, Princeton University, 1993.

[9] Dolan, W. et al., 1993, Automatically Deriving Structured Knowledge Bases From On-Line Dictionaries, ftp://ftp.research.microsoft.com/pub/tr/tr-93-07.ps, Microsoft.

[10] LDOCE (Longman Dictionary of Contemporary English, http://www.ldoceonline.com.

[11] http://www.ai.mit.edu/projects/infolab.

[12] Navigli, R,Word sense disambiguation: A survey. ACM Comput. Surv. 41, 2, Article 10 (February 2009), 69 pages DOI = 10.1145/1459352.1459355 http://doi.acm.org/10.1145/1459352.1459355, 2009.

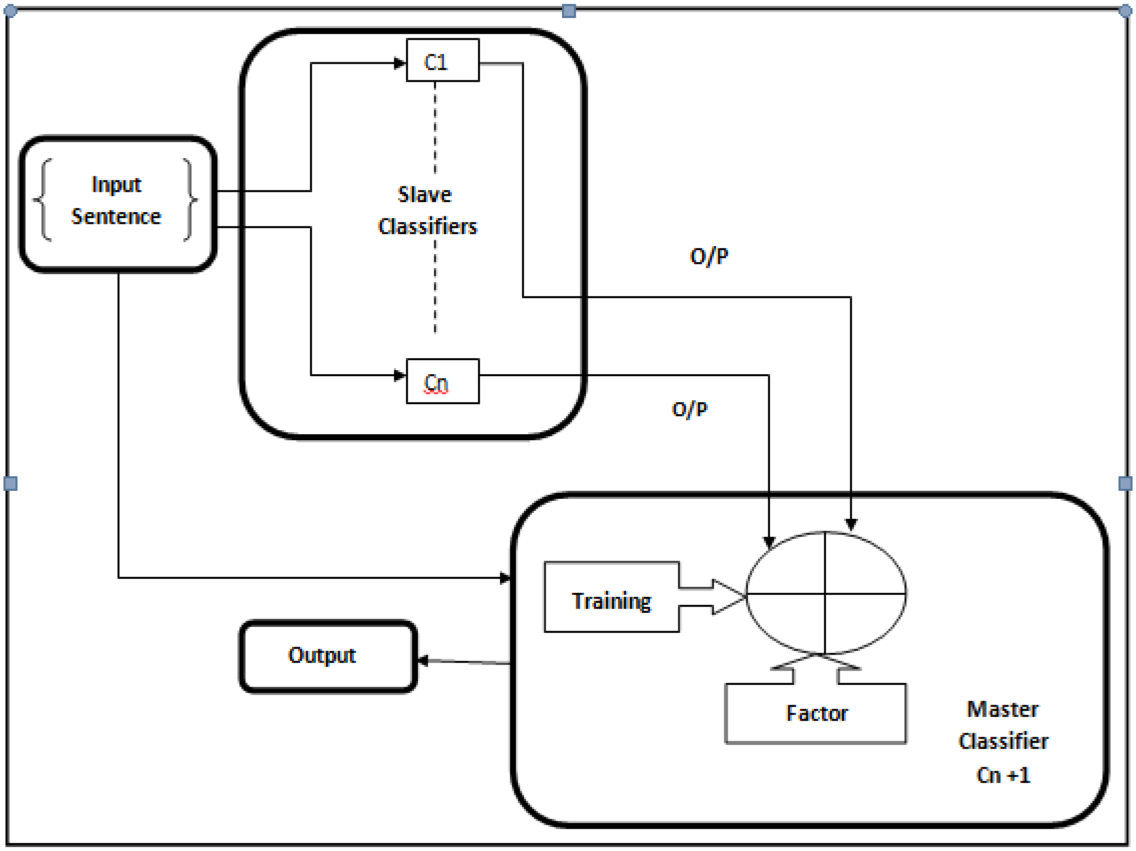

Fig. 1 The Master- Slave voting technique