Development of Hybrid Algorithm for Automatic Extraction of Multiword Expressions from Monolingual and Parallel Corpus of English and Punjabi

Kapil Dev Goyal, Vishal Goyal
Department of Computer Sciecne,
Punjabi University Patiala
{kapildevgoyal,vishal.pup}@gmail.com

Abstract

Identification and extraction of Multiword Expressions (MWEs) is very hard and challenging task in various Natural Language processing applications like Information Retrieval (IR), Information Extraction (IE), Question-Answering systems, Speech Recognition and Synthesis, Text Summarization and Machine Translation (MT). Multiword Expressions are two or more consecutive words but treated as a single word and actual meaning this expression cannot be extracted from meaning of individual word. If any systems recognized this expression as separate words, then results of system will be incorrect. Therefore it is mandatory to identify these expressions to improve the result of the system. In this report, our main focus is to develop an automated tool to extract Multiword Expressions from monolingual and parallel corpus of English and Punjabi. In this tool, Rule based approach, Linguistic approach, statistical approach, and many more approaches were used to identify and extract MWEs from monolingual and parallel corpus of English and Punjabi and achieved more than 90% f-score value in some types of MWEs.

1 Introduction

In this tool, ruled based, linguistic and statistical approaches are used to extract MWEs. Apart from these approaches, Part of Speech tagger tool, Named Entities Recognizer tool, Stemmer, Giza++, etc tools are used. Most of MWEs are generic in nature, it means based on rules and linguistic approach than statistical approach. Mostly ruled based approach is used in Replicated words, which are strong candidate of MWEs in Punjabi Language. The results of linguistic approach in English are better than results of Punjabi, because of the poor performance of Punjabi Part of Speech tagger tool and Punjabi Stemmer. Similarly results of monolingual corpus are better than parallel corpus, because of the lack of proper Punjabi-English dictionary and poor performance of giza++ tool. Therefore results of MWEs are directly depending upon performance of above mentioned tools. As we earlier discussed, most of the NLP applications are highly affected by MWEs. This automatic MWEs tool will help the performance of many NLP applications like Information Retrieval (IR), Information Extraction (IE), Question-Answering systems, Speech Recognition and Synthesis, Text Summarization and Machine Translation (MT). Non-compositional, non-modifiable and non-substitutable are basic features of MWEs. Non Compositional means that meaning of MWEs cannot be predicted from meaning its parts. Non Modifiable means that Multiword Expressions are frozen and they cannot be changed in any way. Non Substitutable means that any parts of Multiword Expression cannot be substituted by one of its synonym without affecting the meaning of an expression.

1.1. Features of MWEs

(Manning & Schutze, 1999) described that non-compositional, non-modifiable and non-substitutable are basic features of MWE.

(1) Non-compositional: It means that MWE cannot be completely translated from the meaning of its parts.

E.g. ਅੱਖਾਂ ਦਾ ਤਾਰਾ) Punjabi)

Transliteration: “Akhān dā tārā”
Gloss: Star of Eyes
Translation: Lovely

E.g. लोहे के चने चबाना (Hindi)
Transliteration: “Lōhē kē chanē chabānā”

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 4–6
Patna, India, December 18 - 21, 2020. ©2019 NLP Association of India (NLPAI)

Gloss: To chew iron gram
Translation: Difficult task

In above examples, actual translations cannot be predicted from their parts, which are completely different from its basic meaning.

(2) Non-modifiable:
Many Multiword Expressions are frozen and they cannot be changed in any way. These types of expressions cannot be modified by grammatical transformations (like by changing Number/ Gender/ Tense, addition of adjective etc).
Eg. In ਰੋਜੀ ਰੋਟੀ) Rōjī rōṭī) cannot be changed in number as ਰੋਜੀ ਰੋਟੀਆਂ) Rōjī rōṭī'ān)

(3) Non-Substitutable:
Any word of Multiword Expression cannot be substituted by one of its synonym without affecting the meaning of an expression.
E.g. ਰੋਜੀ ਰੋਟੀ) Rōjī rōṭī) cannot be written as ਰੋਜੀ ਖਾਣਾ) Rōjī khāṇā) or ਰੋਜ ਰੋਟੀ) Rōj rōṭī)

2 Review of literature

The concept of Multiword Expression is given by (Baldwin & Kim, 2010; Sag et al., 2002) has covered all types of MWEs. Identification and extraction of MWEs is not very old field, but still many of the researchers are working in this field. There has been very limited work done reported on monolingual Punjabi MWEs and extraction of parallel MWEs from English-Punjabi parallel corpus. Most of researcher used statistical method or association measures tools (Evert & Krenn, 2005), linguistic based approaches (Goldman et al., 2001; Vintar & Fišer, 2008), ruled based approaches (M Nandi, 2013), hybrid approaches (Boulaknadel et al., 2008; Jianyong et al., 2009) for extracting MWEs expression from monolingual and parallel corpus. In Indian language, (Rácová, 2013; Singh et al., 2016; Sinha, 2009, 2011) classified reduplicate MWEs which are strong candidates of MWEs.

3 Methodology

This report presents an automatic tool which extracts following Multiword Expressions from English and Punjabi Language.

3.1. Punjabi Multiword Expressions:

3.1.1. Replicated Multiword Expression

Replicated MWEs
Combination with antonyms/gender
Combination with hyponyms
'Walaa' Morpheme

3.1.2. Extracted using Statistical Methods

3.1 3. Extracted using Linguistic Methods

Name Entities
Compound Noun
Conjunct Verbs
Compound Verbs

3.1.4. Extracted using Hybrid Approach

3.2. English Multiword Expressions:

3.2.1. Extracted using Statistical Methods

3.2.2. Extracted using Linguistic Methods

Name Entities
Compound Noun
Conjunct Verbs
Compound Verbs

3.2.3. Extracted using Hybrid Approach

3.3. Punjabi-English Parallel Multiword Expressions:

Replicated Multiword Expression
Extracted using Statistical Methods
Extracted using Linguistic Methods
Extracted using Hybrid Approach

4 Results

MWEs Type	Accuracy	Precision	Recall	F-Score
Rule Based Approach	99.88	94.73	81.81	87.80
Using Linguistic Methods	99.67	33.33	75	46.15
Using Statistical Methods	80.03	52.35	29.29	37.56

In these results, accuracy scores are more than 70%, but precision, recall and F-score values are varied from 30% to 97%. Results of replicated words using rule based approach are relatively better than linguistic approach and statistical approach. Statistical tools measure the association between words, therefore results for statistical methods are relatively less than all above types.

References

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. *Handbook of Natural Language Processing, Second Edition*, 267–292.

Boulaknadel, S., Daille, B., & Aboutajdine, D. (2008). A multi-word term extraction program for Arabic language. *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*.

Evert, S., & Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, *19*(4). https://doi.org/10.1016/j.csl.2005.02.005

Goldman, J.-P., Nerima, L., & Wehrli, E. (2001). Collocation extraction using a syntactic parser. *Proceedings of the ACL Workshop on Collocations*, 61–66.

Jianyong, D., Lijing, T., Feng, G., & Mei, Z. (2009). A hybrid approach to improve bilingual multiword expression extraction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-01307-2_51

M Nandi, R. R. (2013). Rule-based Extraction of Multi-Word Expressions for Elementary Sanskrit. *International Journal of Advanced Research in Computer Science and Software Engineering*.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. https://books.google.com/books?hl=en&lr=&id=3qnuDwAAQBAJ&oi=fnd&pg=PT12&dq=Foundations+of+statistical+natural+language+processing.+MIT+press,+1999.+manning&ots=ysF-mZAwM_&sig=GplFqEroiO9dy1ZGYhebtAhhEdk

Rácová, A. (2013). Reduplication of verbal forms in Bengali. *Asian and African Studies*.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2276*, 1–15. https://doi.org/10.1007/3-540-45715-1_1

Singh, R., Ojha, A. K., & Jha, G. N. (2016). *Classification and Identification of Reduplicated Multi-Word Expressions in Hindi*. *May*.

Sinha, R. M. K. (2009). *Mining complex predicates in Hindi using a parallel Hindi-English corpus*. *August*, 40. https://doi.org/10.3115/1698239.1698247

Sinha, R. M. K. (2011). Stepwise mining of multi-word expressions in Hindi. *Workshop on Multiword Expressions: From Parsing and Gerenraion to Real World(MWE 2011)*, *June*, 110–115. http://dl.acm.org/citation.cfm?id=2021121.2021143

Vintar, Š., & Fišer, D. (2008). Harvesting multi-word expressions from parallel corpora. *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*.