Single-Sentence Compression using XGBoost

Deepak Sahoo, IIIT-Bhubaneswar, Odisha, India
Rakesh Chandra Balabantaray, IIIT-Bhubaneswar, Odisha, India

ABSTRACT

Sentence compression is known as presenting a sentence in a fewer number of words compared to its original one without changing the meaning. Recent works on sentence compression formulates the problem as an integer linear programming problem (ILP) then solves it using an external ILP-solver which suffers from slow running time. In this article, the sentence compression task is formulated as a two-class classification problem and used a gradient boosting technique to solve the problem. Different models are created using two different datasets. The best model is taken for evaluation. The quality of compression is measured using two important quality measures, informativeness and compression rate. This article has achieved 70.2 percent in informativeness and 38.62 percent in compression rate.

KEYWORDS
Classification, Parsing, Sentence Compression, Training Data, XGBOOST

1. INTRODUCTION

In natural language processing sentence compression is an important problem where the aim is to generate a small and grammatically correct version of the source sentence. This is an active research area due to various reasons like increasing penetration of smart phones, Use of Internet through smart phone and increase in on-line shopping through smart phones. This situation demands small but complete informative sentence which fits in small screen of smart phone than large sentences.

Sentence compression is also very important in compression based extractive summarization and abstractive text summarization. Text summarization system itself is a very complex system; if the sentence compression module will be a complex then the overall summarization system will be a very complex. As we are working on hybrid approach to abstractive text summarization therefore we require a simple sentence compression module so that the overall complexity of the summarization system will be less and performance of the system in terms of execution time and quality of summary should increase.

Sentence compression can be of two types of single-sentence compression and multi-sentence compression, research on both the type share two common properties either they depend on syntax or they are supervised. In single sentence compression, the sentence is represented in fewer words without compromising the original meaning of the source sentence where as in multi-sentence compression e.g. the work by K. Ganesan, Zhai & Han (2010) where more than one sentence that are highly related (e.g. +ve/-ve view regarding a product or two sentences having transition or anaphoric relationship and some common words) are joined to produce a new sentence whose size (number of words) is less than the combined size (number of words) of original sentences.

DOI: 10.4018/IJIRR.2019070101

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
We worked on single-sentence compression where the new generated sentence is represented in less number of words compared to original sentence without compromising the meaning of original sentence. Let the sentence \( S_i \) belongs to a document \( D \). \( S_i = W_{i1}...W_{in} \), \( i = 1, 2...n \); \( n \) = Total number of words in the sentences. Let after the compression we get the compressed sentence \( CS_i = W_{m1}...W_{mj} \), \( i = 1, 2...m \); \( m \) = Total number of words in compressed sentence and \( m \leq n \).

Example:
Original Sentence: Five people have been taken to hospital with minor injuries following a crash on the A17 near Seafor this morning.
Compressed Sentence: Five people have been taken to hospital with minor injuries following a crash on the A17 near Seafor

In this case \( m<n \). Here \( n = 20 \), \( m = 18 \);

We tried to solve the problem using machine learning approach. The paper is structured as follows: Related word is presented in section 2. Detailed methodology is discussed in section 3. Evaluation of the system is discussed in section 4, Research implications and Limitations given in section 5 followed by the conclusion.

2. RELATED WORK

Dozens of systems have been introduced for sentence compression and most of them are deletion based that heavily relies on syntactic information to minimize grammatical error in the output but that leads to a very complex system. Generally the sentence compression system used as a module in text summarization system so it is required that the compression system should be simple otherwise the overall complexity of the system will be very high.

Different sentence compression systems have been proposed by different researchers since its very first set approach by Grefenstette (1998), Knight & Marcu (2000, 2002), Jing & McKeown (2000). Parsers play an important role in NLP task. Some sentence compression methods highly depend on parsers to identify syntactic information for sentence compression task Clarke & Lapata (2006), McDonald (2006), Toutanova, Brockett, Gamon, Jagarlamudi, Suzuki and Vanderwende (2007), Nomoto (2009), but systems based on syntactic information is not robust.

To produce grammatically correct compressed sentence some methods play with the syntactic structure of the tree Galley & McKeown (2007), Cohn & Lapata (2009), Filippova & Strube (2008a), Galanis & Androustopoulos (2010), Wang, Raghavan, Castelli, Florian & Cardie (2013) use techniques that modify/rectify the syntactic trees.

Some researchers are used rule-based approach Dorr, Zajic & Schwartz (2003) and language-based model Horii, Furui, Malkin, Yu & Waibel (2003), Clarke & Lapata (2008), for sentence compression. The drawback of the rule-based system is that rules may not cover all types of sentences.

Sentence compression and extractive approach for summarization are coupled by Liu & Liu (2013) to generate summary for the spoken documents; Learning-based model designed by Wang (2013) that rely on parse trees and its module of query based multi-document summarization. Prior work on sentence compression heavily rely on generic sentence compression approaches McDonald, (2006); Nomoto (2007); Clarke & Lapata, (2008); Thadani & McKeown (2013) to compress the sentences in a document, yet a generic compression system may not be the best fit for the summarization purpose. combine sentence compression with summarization system and frame it as an integer linear programming (ILP) Martins, Smith (2009), Berg-Kirkpatrick, Gillick & Klein (2011). Thadani & McKeown (2013) then solve the problem with an external IP-solver. But the issue with integer linear programming approach is that they suffer from slow run time. The frame work proposed by Woodsend & Lapata (2012), Almeida & Martins (2013) based on integer linear programming but they have improved running time compared to previous ILP-based systems.

Filippova & Alfonseca (2015) proposed a supervised model that based on local deletion decision in association with a recursive procedure to get most probable compression at every node in the
Related Content

On Knowledge-Enhanced Document Clustering
[www.igi-global.com/article/knowledge-enhanced-document-clustering/78315?camid=4v1a](http://www.igi-global.com/article/knowledge-enhanced-document-clustering/78315?camid=4v1a)

Multimodal Information Systems
[www.igi-global.com/chapter/multimodal-information-systems/197709?camid=4v1a](http://www.igi-global.com/chapter/multimodal-information-systems/197709?camid=4v1a)

Systems Administration in Ontology-Based Applications: The Case of Citizen Relationship Management
[www.igi-global.com/chapter/systems-administration-in-ontology-based-applications/198607?camid=4v1a](http://www.igi-global.com/chapter/systems-administration-in-ontology-based-applications/198607?camid=4v1a)
Compressing and Vague Querying (XCVQ) Design
www.igi-global.com/chapter/compressing-vague-querying-xcvq-design/69133?camid=4v1a