CMPUT 650 Project Proposal: SemEval2024 Task 1 - Semantic Textual Relatedness using Concept Overlap

Jai Riley
Department of Computing Science
University of Alberta
Edmonton, CA
jrbuhr@ualberta.ca

Mohammad Tavakoli
Department of Computing Science
University of Alberta
Edmonton, CA
tavakol5@ualberta.ca

Abstract

Semantic Textual Relatedness (STR) refers to the extent of meaning overlap between two or more pieces of text. It is a crucial concept in natural language processing (NLP) and computational linguistics, aiming to quantify the semantic similarity or closeness between textual units. The goal of SemEval2024 Task 1 is to develop computational models that can understand and measure the similarity of meaning between two sentences in languages such as English, Spanish and different Asian and African languages. Unlike traditional methods that rely on lexical and syntactic features, our approach focuses on using Word Sense Disambiguation (WSD) techniques to capture the senses of each word and compare their underlying concepts. By measuring the overlap of these identified concepts between two texts, our approach aims to provide a more nuanced and contextually relevant assessment of semantic relatedness.

1 Introduction

STR has emerged as a pivotal area of study within natural language processing, focusing on quantifying the degree of meaning overlap between linguistic units such as words, phrases, sentences, and other components of language (Mohammad, 2008; Mohammad and Hirst, 2012). As language inherently thrives on context and semantics, accurately measuring the relatedness of textual elements is critical for applications ranging from information retrieval to sentiment analysis. Additionally, STR contributes to machine translation by preserving meaning across languages. Paraphrase detection, understanding context, and improving human-computer interaction are other key areas where STR proves essential. In this project, we introduces an approach named concept overlap as a solution to address the inherent complexities in measuring semantic relatedness. Unlike traditional methods, concept overlap relies on extracting underlying concepts within the text using a WSD system and comparing there overlap. The approach aims to provide a more nuanced and contextually relevant assessment of semantic relatedness.

2 Related Work

In the past, researchers have utilized the concept of semantic relatedness between words or concepts in various applications, such as text summarization (Barzilay and Elhadad, 2002), text retrieval (Stokoe et al., 2003), and WSD tasks (Patwardhan et al., 2003). These measures are broadly categorized into dictionary-based, corpus-based, and hybrid approaches.

Within dictionary-based measures, the measure of Agirre and Rigau (1997) was among the earliest developed to calculate semantic relatedness among two or more concepts within a set. Their approach considers the density and depth of concepts in the set and the length of the shortest path connecting them, assuming equal importance for all edges in the path.

One corpus-based measure, introduced by Leacock et al. (1998), assesses the semantic similarity between a pair of concepts by considering the length of the shortest path connecting them (measured in the number of nodes in the path) and the maximum depth of the taxonomy.

In terms of hybrid approaches, Resnik (1999) proposed a measure for pairs of concepts that relies on the Information Content (IC) of the most profound concept capable of encompassing both concepts, known as the least common subsumer.

3 Methodology

This project aims to tackle the issue of STR by employing the concept overlap approach. Our objective is to calculate the intersection of concepts used in two pieces of text. The methodology involves extracting and comparing the underlying concepts present in each text, aiming to quantify the degree of overlap in their meaning. This approach holds

Language	Pearson Corr.
English	0.6158
Spanish	0.5572
Algerian Arabic	0.3055
Moroccan Arabic	0.4485
Amharic	0.5677
Hausa	0.3714
Kinyarwanda	0.0402
Marathi	0.6518
Telugu	0.7062

Table 1: Preliminary Results with AMuSE-WSD for concept overlap.

promise in addressing the nuances inherent in STR, offering a more nuanced and contextually relevant assessment of the semantic relatedness between linguistic elements.

3.1 Experimental Setup

The original WSD system chosen for our idea is AMuSE-WSD (Orlando et al., 2021) due to its ease of use. Unfortunately, we were only able to access the online version as our application to gain access to the offline Docker version of AMuSE-WSD was denied. To the best of our knowledge, the only overlap in the languages used to train AMuSE-WSD with the languages given for the SemEval2024 Task 1 are English and Spanish. Our next step for this project is to try out different WSD systems that may even be able to handle the untranslated versions of the given datasets to improve upon our preliminary results.

3.2 Datasets

Datasets have been provided through SemEval Task 1 for 14 languages: Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Punjabi, Spanish, and Telugu. For the original submission, we used English translations of the given datasets as only two languages were supported by the online version of AMuSE-WSD.

4 Evaluation and Preliminary Results

Our original approach calculated a similarity score between two sentences by taking the intersection of the overlapping concepts divided by the union of all words in the sentences that were given a sense. In future work we hope to calculate similarity score by summing the similarity of all concepts and dividing it by the number of words given senses. In order to measure the performance of our approach, the Pearson correlation can be calculated between our predicted similarity score and the gold STR score. Preliminary findings are displayed in Table 1.

References

Eneko Agirre and German Rigau. 1997. A proposal for word sense disambiguation using conceptual distance. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 161–172.

Regina Barzilay and Noemie Elhadad. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Claudia Leacock, Martin Chodorow, and George A Miller. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

Saif Mohammad. 2008. *Measuring semantic distance using distributional profiles of concepts*. University of Toronto.

Saif M Mohammad and Graeme Hirst. 2012. Distributional measures of semantic distance: A survey. *arXiv preprint arXiv:1203.1858*.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. Amuse-wsd: An all-in-one multilingual system for easy word sense disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, page 298–307, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Computational Linguistics and Intelligent Text Processing: 4th International Conference, CICLing 2003 Mexico City, Mexico, February 16–22, 2003 Proceedings 4*, pages 241–257. Springer.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 11:95–130.

Christopher Stokoe, Michael P Oakes, and John Tait. 2003. Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 159–166.