Semantic Textual Similarity Evaluation for Spanish Sentences

A Major Qualifying Project
Submitted to the faculty of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the Degree
of Bachelor of Science

April 27, 2017

Fiona Heaney
Matthew Zielonko

Advisor
Gabor Sarkozy

Abstract

This project sought to enhance the natural language processing research of the MTA SZTAKI institute in Budapest, Hungary, by extending their semantic textual similarity system to evaluate Spanish sentences. Language analysis resources were collected to generate a working system to analyze similarities between Spanish sentence pairs. This system was based on that which the institute had previously developed for English. The final system was tested against large data sets of sentence pairs, and compared to a Gold Standard of scores created by human linguists, with the goal of having a high correlation between the two data sets.

Acknowledgements

We would like to extend our sincerest gratitude to those who made this project possible and successful:

- András Kornai for his oversight and feedback throughout the project
- Judit Ács, Dávid Nemskey, and Gábor Recski for their advice and guidance
- MTA SZTAKI for providing us with project facilities and resources
- Gábor Sárközy for advising us and teaching us the value of Hungarian culture
- Worcester Polytechnic Institute for providing us with this opportunity

Contents

Tables

Figures

1.0 – Introduction

Alan Turing posed the question "Can machines think?" in 1950 [9] which has driven artificial intelligence and human computer interaction development. One field that relates to these human computer interactions is that of Natural Language Processing (NLP). The NLP field focuses on designing and building computer systems that can analyze, understand, and generate natural human languages [9]. Every day, over 2.5 quintillion bytes of data are generated [20], and NLP allows researchers to analyze and interpret information from sources such as social media, new publications, and music. Humankind is rapidly moving from the Age of Content to the Age of Context, and NLP is vital in order to extrapolate those contexts.

Through the creation of evaluation systems, computer scientists are able to bridge the gap between what people say and how the things they say relate to the world around them [28 pg 3-5]. The usage of techniques such as pattern matching [18 pg 7-10], syntactically driven parsing [18 pg 10-16], semantic grammars [18. pg 17-20], and case-frame instantiation [18 pg 20-24] enables researchers to create computer understandable representations of words. Methods for the evaluation of these computer understandable representations have evolved greatly in the last decade, partially due to the founding of *Semantic Evaluation* (*SemEval*) [1] challenges in 2012. The tasks proposed by organizers seek to further progress in assessing *Semantic Textual Similarity* (*STS*) [19] between two texts, such as sentences or Twitter data.

Our project contributes directly to the availability of Spanish STS tools and resources through the creation of Spanish gold standard evaluation sets and modification to a STS evaluation framework, originally published in 2013 by Han et al. [17]. The goal of our project was to create and test sets of hyperparameters for the modified Han et al. framework in order to deliver more accurate STS evaluation of the 2014 SemEval Task 10 datasets [2] than original participants. We were able to complete this task by evaluating three sets of 300 dimension word vectors and adding modifications to an English STS evaluation framework to include Spanish dependency files, thus allowing us to create sets of hyperparameters tailored to Spanish STS performance.

2.0 – Background

The field of *Natural Language Processing*, or *NLP*, is a combination of the computer science, artificial intelligence, and computational linguistics disciplines. It aims to bridge the gap between human created languages and computational processing for activities such as machine translation, natural language generation, creation of human-computer dialog systems, and tasks involving human speech or writing translation into machine understandable representations. Throughout the NLP field, there is a need for a variety of criteria for assessing machine performance, as well as tools that aid researchers in rapidly processing large language corpora.

2.1) Semantic Textual Similarity

Semantic Textual Similarity, or *STS*, measures the degree of semantic equivalence between two texts [19]. In recent years, the field of Natural Language Processing presented a need for computational methods to determine equivalence between sentences or short texts [2]. Examining textual similarity improved machine classification of sentence meaning and sentence relationships over precursor methods, such as identifying part of speech (POS) or sentence structure patterns, and instead focused on the content conveyed in the sentence or passage. STS is powerful and has implications in activities such as web page retrieval, machine translation, and general question answering, among others [6] .

There are three major types of semantic textual similarity approaches: vector space, alignment, and machine learning. With vector space approaches, texts are represented as vectors and a vector similarity (eg. cosine similarity or scalar dot product) between the two vectors is used to compute similarity [6]. Cosine similarity is most common because the length of the given vectors can be normalized in order to eliminate biases based on sentence or document length, instead looking at the vector similarity in terms of their relative position to one another. Alignment approaches take the words and phrases in each text and align them against one another and the coverage of the resultant is used as a similarity measure [6]. The last approach, machine learning, relies on the usage of training data to perform supervised learning [6]. Machine learning performance is heavily reliant on the amount of training data available, such that more training data allows for better trained machine performance.

2.2) Language Processing Components

In order for researchers to collect and analyze the abundance of words in a given language, certain resources and tools have been created. For the purposes of assessing semantic textual similarity, a language corpus, embedding set, and similarity set are vital. Each component is used for language parsing, learning, or transformation in order to translate languages into machine understandable representations.

2.2.1) Corpora

A corpus, in the linguistic sense, is a collection of written or spoken texts that offers an extensive description of a language. According to the Oxford English Dictionary, a corpus "provides the evidence of how language is used in real situations, from which lexicographers can write accurate and meaningful dictionary entries" [33].

Language corpora are most often considered to be open, meaning that the data that they provide is not a complete representation of the language. In other words, because each language is so complex and intricate, there may be words missing from the corpus [33]. Other corpora however, for example, historical corpora, do claim to represent a completed dataset, making them closed. In the computer science and information processing sense, a computer-processable corpus can keep track of all instances of words and phrases, which can prove to be valuable for language analysis [33].

2.2.2) Embeddings

An embedding is a method or function used to map text to vector representations. Within the context of natural language processing, the main embedding technique is word embedding, in which words in a collection of text are mapped to vectors. There are, however, other types of embeddings used for natural language processing. For instance, phrase embedding, which performs the same set of tasks as word embedding, but with sets of words as opposed to individual ones. These techniques have been known to enhance and improve the performance of language analysis and parsing tasks.

Vector Space Models (VSM) embed words into a vector space, such that semantically similar words are embedded near each other [16]. There are two categories of Vector Space Models: count-based and predictive models. Google's TensorFlow describes the differences between the two categories such that "count-based methods compute the statistics of how often some word co-occurs with its neighbor words in a large text corpus, and then map these count-statistics down to a small, dense vector for each word. Predictive models directly try to predict a word from its neighbors in terms of learned small, dense embedding vectors (considered parameters of the model)" [16].

2.2.3) Similarity Set

For the purposes of evaluating STS performance, the NLP field has produced certain gold standard sets of human graded similarity pairs, called similarity sets. These data sets are typically small samples of a given language and consist of pairs of words whose part of speech (POS), concreteness, and similarity to one another are documented [12]. By comparing experimental results to a given gold standard, researchers are able to compare the effectiveness of their proposed system compared to human classification.

2.3) Language Processing Tools

When analyzing a language, it is also vital to have the right set of tools to process the corpus. Firstly, there must be a way to take a sentence and split it into smaller parts while still maintaining the structural integrity of the sentence. Secondly, there needs to be a way to take each of the individual words of a sentence, and analyze the usage within the

sentence to figure out what context the word is being used in. For these two tasks there are dependency parsers, and WordNets, respectively.

2.3.1) Dependency Parser

A dependency parser is a tool that analyzes text and creates graphs and trees of out sentences. The words in the sentence are nodes, with action verbs and words that are vital to the overall structure of the sentence serving as root and parent nodes, and the children are the words that are the subjects and objects of the sentence. Each of the edges within the tree are labeled with the relationship between the words [14].

A similar form of graphically analyzing sentences is constituency parsing, where text is divided into and analyzed as phrases, or combinations of words. In this case, the sentence serves as the root of the tree and the subsequent children are the individual words and phrases in the sentence in the order they appear [14]. Figure 1 below is an example of how the same simple sentence, "John loves Mary", might be represented across multiple parsers.

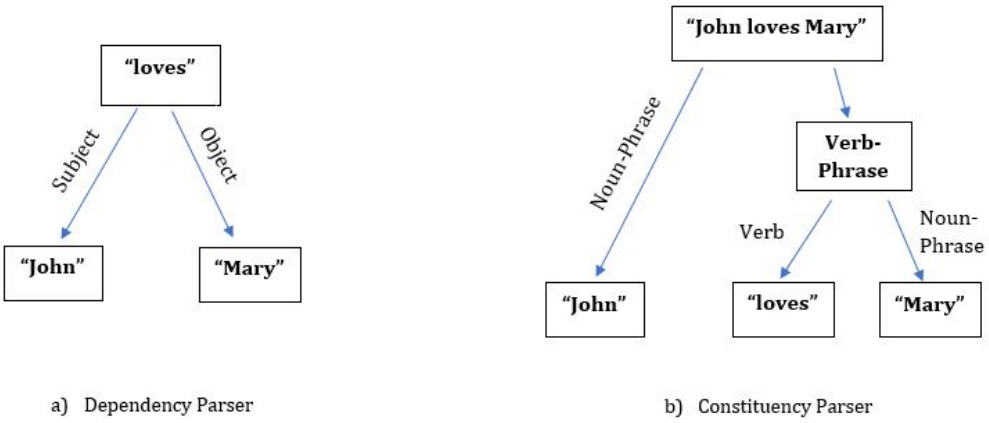

Figure 1: Different types of parsers analyzing the same sentence. a) A dependency parser finds the action verb, and uses it as the root of the tree. b) The constituency parser creates a verb-phrase from "loves Mary", and breaks down the sentence sequentially.

2.3.2) WordNet

A WordNet is a large database of words that are all interconnected through a graphical web of linguistic and syntactical relationships. The most developed WordNet in the NLP field is currently Princeton University's WordNet. Princeton University's team formally defines a WordNet as "a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept" [38]. The Princeton WordNet contains many additional features to help further analyze the vocabulary that it has collected. This includes definitions for each of the synsets associated with every word and synonyms of each word within a synset, known as lemmas [38].

Each synset is interconnected with semantic and lexical relations. In this way, WordNets resemble a thesaurus since it groups words based on their meanings with the exception that words are labeled based on semantic relations from word to word, whereas a thesaurus does not follow any pattern besides meaning similarity [38].

2.4) SemEval Challenges

The SemEval, or Semantic Evaluation, challenges began in 2012 with a focus on the evaluation of semantic analysis systems in hope of furthering progress towards creating systems capable of analyzing diverse phenomena in text [1]. The challenges were first created by the ACL-SIGLEX [32] and ACL-SIGEM for a joint conference on lexical and computational semantics. The challenges have grown more complex from year to year, and most recently included five tracks: textual similarity, sentiment analysis, semantic parsing,

semantic analysis, and semantic taxonomy [4]. Each challenge is posed with accompanying training and/or test datasets with the target audiences being individual researchers or teams.

2.4.1) SemEval 2014 - Task 10

In 2014, there were multiple challenges put forth for a variety of languages but task number ten specifically dealt with the assessing STS in regards to the Spanish language. The ultimate goal of the subtask was to "enable the evaluation of semantic textual similarity systems for Spanish" [2]. Given two sentences in the format `s1 <tab> s2` participants were required to compute how similar the sentences were and return a similarity score (0-4 where 4 is a set of paraphrases and 0 means no relation), as well as an optional confidence score [2]. The Gold Standard for sentence scoring can be found in Table 1.

The test datasets were comprised of sentence pairs derived from Wikipedia articles and Spanish news articles. Wikipedia sentences were collected from a 2013 parsing of the Spanish version of Wikipedia [2], while the news sentences were extracted from Spanish news publications written in 2014 using mining on the Google News Spanish service [2]. From these two datasets, a development dataset of 65 annotated sentence pairs was created, along with two test datasets containing 324 and 480 sentence pairs. No training data was given for this task, however participants were allowed to use the development dataset for this purpose [2].

Score	**Spanish Sentences**	**English Translations** *(Given here for context)*
(4) The two sentences are completely equivalent, as they mean the same thing	• El pájaro se está bañando en el lavabo. • El pájaro se está lavando en el aguamanil.	• The bird is bathing in the sink. • Birdie is washing itself in the water basin.
(3) The two sentences are mostly equivalent, but some details differ.	• John dijo que él es considerado como testigo, y no como sospechoso. • “Él ya no es un sospechoso," John dijo.	• John said he is considered a witness but not a suspect. • "He is not a suspect anymore." John said.
(2) The two sentences are roughly equivalent, but some important information differs/missing.	• Ellos volaron del nido en grupos. • Volaron hacia el nido juntos	• They flew out of the nest in groups. • They flew into the nest together.
(1) The two sentences are not equivalent, but are on the same topic.	• La mujer está tocando el violín. • La joven disfruta escuchar la guitarra.	• The woman is playing the violin. • The young lady enjoys listening to the guitar.
(0) The two sentences are on different topics.	• Al amanecer, Juan se fue a montar a caballo con un grupo de amigos. • La salida del sol al amanecer es una magnífica vista que puede presenciar si usted se despierta lo suficientemente temprano para verla.	• John went horseback riding at dawn with a whole group of friends. • Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

Table 1: SemEval 2014 Task 10 Example Scoring. English sentences not provided in dataset but given here for reader understanding.

3.0 – Resources

In order to complete this research project, we compiled a set of resources that satisfy all of the requirements laid out in the sections above. These resources proved helpful for script writing and data analysis, as they contributed to achieve the goal of analysing Spanish words in context through similarity. We were able to utilize a 1.5 billion word corpus, three sets of embeddings, a word net, a dependency parser, and a word similarity set that was translated from English into Spanish.

3.1) Corpus

The Spanish Corpus used for this study was taken from the research of Cristian Cardellino at the University of North Carolina [13]. Cardellino compiled and analyzed a set of one and a half billion Spanish words from various web and text resources, and has provided his corpus, and resulting word vectors online [13].

The corpus is comprised of multiple free novels available in Spanish online, such as Jane Austen's “Sense and Sensibility” and Lewis Carrol’s “Alice in Wonderland”, as well as documents from the European parliament. Spanish portions of pre-existing corpora, namely SenSem and the Ancora Corpus, were also included in the data set. Finally, Cardellino used a Wikipedia Extractor to perform a dump of Spanish words from Wikipedia articles. Cardellino noted the cleaning he performed on the Corpus by adding, “...we proceed to replace all non-alphanumeric characters with whitespaces. All numbers with the token ‘DIGITO’ and all the multiple whitespaces with only one whitespace” [13]. All of

these resources were compiled together to assure that we used the most complete dataset available throughout our testing.

3.2) Embeddings

In traditional natural language processing systems, words are treated as discrete atomic symbols (i.e. 'cat' could be represented as id-537), but the symbols are arbitrary, and lack useful information about relationships to other words [16]. For instance, if the word 'cat' is preceded by 'dog', valuable shared data such as 'four legged', 'pet', or 'animal' is not conveyed. By using vectors instead of discrete symbols, words with similarities can be grouped together in the vector space.

3.2.1) Spanish Billion Word Embeddings

The set of embeddings associated with the billion word corpus from Cristian Cardellino at the University of North Carolina came as a downloadable set of word-vector pairs in a .txt file. Cardellino also provided some basic analysis of the embeddings on his website. The pairs were developed using a word2vec algorithm, which uses a skip-gram model to map words to their corresponding vectors. Jianpeng Jang and his team at the University of Oxford described the concept of skips-grams as follows:

> *"The essence of a neural language model is to discover salient word representations that best interpret the co-occurrence relations among words. This is instantiated in the Skip-gram as updating the word representations so that the model can predict the likely contexts of a target word. An assumption made here is that words that appear often in similar contexts tend to have similar meanings, and hence should be assigned similar representations." [15]*

Cardellino also included his hyperparameters, stating that within the corpus, words had to come up at least five times to be considered. “Noise words”, such as “de” “en”, and “y” were limited to twenty, and the dimension of the final vectors were all of size three hundred [13]. He also removed words that he classified as “ambiguous”, such as various currency names that all mapped to the English word “crown”, cutting the sample size by about a quarter [13].

3.2.2) Facebook Spanish Embeddings

The set of Facebook Spanish Embeddings was created using Facebook’s fast Text library (Source: Bag of Tricks for Efficient Text Classification) and is composed of pre-trained word vectors of dimension three hundred obtained using Wikipedia data. Facebook’s contributors describe fastText as “a library for efficient learning of word representations and sentence classification” [22]. The vectors were obtained using a skip-gram model [11] with default parameters that are similar in approach to the methods outlined in Section 2.2.1. These embeddings were downloaded to MTA SZTAKI’s server and were recommended for use by Professor Kornai due to their large coverage of the Spanish Wikipedia.

3.2.3) GloVe Parsed Embeddings

GloVe, which stands for Global Vector, is a vocabulary learning algorithm developed at Stanford University by Jeffery Pennington, Richard Socher, and Christopher Manning [35]. According to the developers, “GloVe is essentially a log-bilinear model with a

weighted least-squares objective. The main intuition underlying the model is the simple observation that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning" [35].

GloVe is designed in such a way that pairs of words are related to one another by concept, and are mapped to one another by representing the words as vectors [34]. Comparing certain pairs of words in this way can be largely influenced by predefined pairs of words that have matching concepts [35]. For example, in Figure 2 below, the GloVe developers demonstrate how the relation between "man" and "woman" might be mapped into a vector space. The distance and angle of the resulting vector is directly influenced by vectors such as "king-queen" and "sister-brother", due to the unifying concept of gender. For this reason, the distance between "man" and woman" is similar to the distance between the other vectors that are unified by gender.

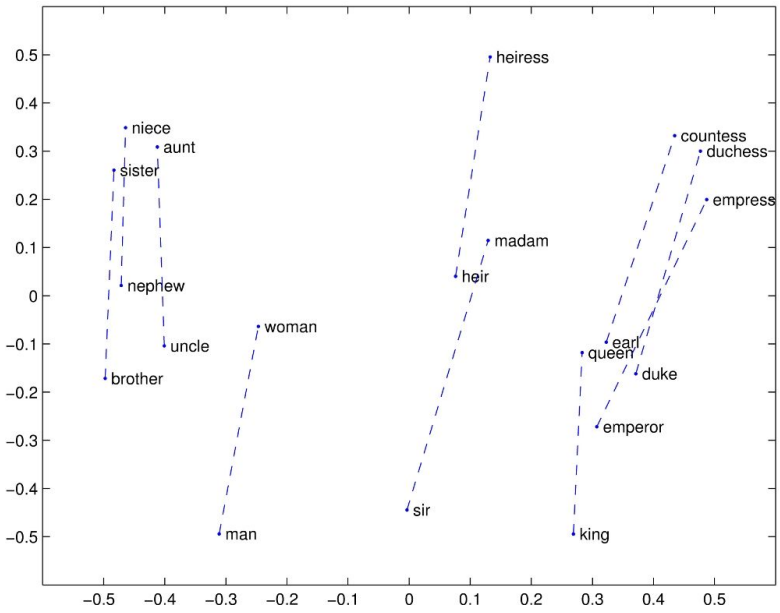

Figure 2: A visual representation of a GloVe mapping system using the relation "man-woman" and other vectors that can be mapped similarly.

For the purposes of this study, we passed the same corpus associated with the word2vec algorithm into the GloVe tool using a short Python script, which outputs an embedding set of Spanish words mapped to vectors. We opted to keep the hyperparameters consistent with those of Crisitan Cadellino, so word frequency is set to a minimum of five, and all resulting vectors are located in a three hundred dimensional space. Given the similarities between the hyperparameters and data set, we aimed to achieve a similar vector set to the one provided by Professor Cadellino.

3.3) Similarity Set

SimLex-999 is a gold standard English similarity set used for model evaluation pertaining to word and concept learning. SimLex-999 provides a way to measure how well

models capture similarity, rather than relatedness or association. [19]. The data set is a tab separated plaintext file with the following properties

Property	**Description**
word1	The first concept of the pair
word2	The second concept of the pair.
POS	The majority part of speech of the concept words. Only pairs of matching POS are included in SimLex-999
SimLex-999	The SimLex-999 similarity rating. Average annotator scores are (linearly) mapped from the range [0,6] to the range [0,10] to match other datasets such as WordSim-999
conc(w1)	The concreteness rating of word1 on a scale of 1-7. Taken from the University of South Florida Free Association Norms database.
conc(w2)	The concreteness rating of word1 on a scale of 1-7. Taken from the University of South Florida Free Association Norms database.
conc(Q)	The quartile the pair occupies based on the two concreteness ratings. Used for some analyses in the above paper
Assoc(USF)	The strength of free association from word1 to word2. Values are taken from the University of South Florida Free Association Dataset
SimAssoc333	Binary indicator of whether the pair is one of the 333 most associated in the dataset (according to

	Assoc(USF)). This subset of SimLex-999 is often the hardest for computational models to capture because the noise from high association can confound the similarity rating.
SD(SimLex)	The standard deviation of annotator scores when rating this pair. Low values indicate good agreement between the 15+ annotators on the similarity value SimLex-999. Higher scores indicate less certainty.

Table 2: SimLex-999 Dataset Description. Each descriptor is one of the columns for a given row entry (Source: SimLex-999 Readme)

It is worth noting how the concreteness measures were calculated, as they are one of the more abstract properties of the SimLex-999 data set. The creators of the SimLex-999 set formally define concreteness as "the extent to which a concept has directly perceptible physical referent" [19]. A conceptual phrase such as "truth" will have a lower value for concreteness than an object which is more easily relatable such as "house" or "man".

3.4) Baseline Framework

The top performer of the 2013 SemEval shared task was the UMBC _EQUITY-CORE team, whose framework [17] will was adopted and modified for the sake of this project by the MathLingBudapest team from SZTAKI. The UMBC_EQUITY-CORE team came up with an *align-and-penalize* approach, as well as two vector regression models that include general and domain specific features. The power of their approaches is drawn from a combination of the Latent Semantic Analysis (LSA) and WordNet which will be substituted for the purposes of this project (see Sections 2.4.1 and 2.4.2). Specifically, this project will adopt the *align-and-penalize* approach with modified hyperparameters.

3.4.1) WordNet

Due to the fact that the Spanish language has not been documented and analyzed as greatly as English, the open source WordNets that are currently available are not as developed as the Princeton WordNet. For the purposes of this project, we used MultiWordNet, which is a multilingual lexical database consisting of Italian, Spanish, Portuguese, Hebrew, Romanian, and Latin WordNets [23]. Although the Italian WordNet strictly aligns with the Princeton WordNet, the Spanish portion of the lexical database was made possible through the TALP Group at the Universitat Politecnica de Catalunya (Spain), and therefore has not been guaranteed to hold an equally high standard or quantity of content [23].

3.4.2) Dependency Parser

MaltParser is a dependency parser that utilizes transition steps, meaning that a central automation performs shift-reduce operations. These operations involve creating a parsing tree on a given sentence starting at the youngest set of leaves, and working towards the root, which is the start of the grammar. The transitions between these operations manage the sentence word by word to calculate dependency. Munsta Padro of the Institute of Informatics in Rio Grande, Brazil has performed multiple studies that test the limits of dependency parsing on Spanish corpora. He notes in one of his papers that MaltParser "...was one of the best parsers in the CoNLL Shared Tasks in 2006 and 2007" [34]. For this reason, we decided MaltParser was the best option for dependency parsing for our project.

3.4.3) Part of Speech Tagger

A part of speech tagger, in its simplest form, takes in a word and returns a label for the part of speech the word represents (eg. noun, verb, adjective, etc.). Having sentences in which the POS for each word is given makes it possible to align two sentences based on their matching POS components. TreeTagger [36] is a tool for labeling words in a sentence with their POS and lemma information. Developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart, TreeTagger has the ability to identify POS components in German, English, French, Spanish, Dutch, Italian, Bulgarian, Russian, Portuguese, Galician, Chinese, Swahili, Slovak, Slovenian, Latin, Estonian, Polish, Romanian, Czech, and Coptic.

3.5) MTA SZTAKI - 2015 SemEval Submission

Rather than work directly from the 2013 Han et al. framework, we modified an instance of the framework that our sponsoring agency had created for the 2015 SemEval Task 1 and Task 2 [8][39] submitted under the team name MathLingBudapest. They focused predominantly on Task 1, but made their code accept different configuration files, such that the only thing necessary to run Task 2 was a separate config file.

3.5.1) SemEval 2015 - Task 1

Task 1 of the 2015 SemEval challenge, entitled *Paraphrase and Semantic Similarity in Twitter* [8] was a part of the STS track, but rather than looking strictly at similarity, focused on predicting whether two sentences imply the same meaning. The team was

provided with 18,000 instances of tweets that had been annotated by Amazon Mechanical Turk for training, 1,000 tweets for testing, and two baselines. Ultimately, the team finished in fifteenth out of eighteen teams for Task 1 with their modified Han et al framework.

3.5.2) SemEval 2015 - Task 2

Task 2 of the 2015 SemEval challenge [6] was the continuation of the original STS task set out in 2012, as discussed in Section 2.1. The task was offered for the Spanish and English language, with MathLingBudapest focusing on the English language. The major difference in the 2015 task over previous years was the more specific scoring of sentences, moving from a (0-4) scale to a (0-5) scale, as well as the addition of an optional confidence score. The datasets provided in English included new headlines, image captions, student answers, answers to questions on public forums and sentences expressing committed belief. The team finished seventh out of thirteen teams for this task.

3.5.3) Align and Penalize Adaptations

The MathLingBudapest team sought to make their code highly modular and easy to create configurations for. As a result, their system takes in a user defined configuration file consisting of hyperparameters, an embedding set, and a mode for assessing final score. The different modes available are average, max, and min. As their names indicate, each mode returns the average, maximum, or minimum score for a given pair. Additionally, there are five penalty flags that can be set in the configuration file as well. These penalties are as follows:

- Sim_too_low - penalizes sentences with a similarity confidence less than 0.05

- Penalize_antonyms - penalizes sentences that contain antonyms by 0.5 for each instance of an antonym pair
- Penalize_named_entities - penalizes sentences where a named entity is identified one or both sentences
- Penalize_questions - penalizes sentences that start with words defined in a question starter resource file
- Penalize_verb_tense - penalizes sentences with differing verb tenses

Each penalty is set using boolean values and is False by default but can easily be changed in each configuration file.

4.0 – Methodology

For this research project, we took an extensive Spanish corpus, and used it to create an embedding set which allowed us to analyse the language based on concepts, phrases, and similarities. We passed the corpus through a word analyzation tool called GloVe, which stands for Global Vectors. The output of this program was an embedding set of Spanish words that we compared with two other sets that were provided. Once we decided which of the three embedding sets had the best STS performance evaluation potential, we modified the MathLingBudapest framework using the resources listed in Section two and the selected embedding set.

4.1) SimLex-999 Translation

The first major milestone in our embedding analysis task was obtaining a gold standard of word pairings and similarity scores. Since Simlex-999 was not yet available in Spanish, we utilized the English similarity set and translated the words using a Spanish dictionary. Using these translated words, along with the similarity scores from the original English set, we were able to create our own, unique Spanish Simlex-999 dataset.

While translating the data, we took note of a few idiosyncrasies in the Spanish language that might have hindered the performance of our embedding analysis. Primarily, many pairs of synonyms in the English language map to the same word in Spanish. For example, the words “smart” and “intelligent” in English both map to the Spanish word “inteligente”. For that reason, we also developed a second gold standard set, which we

called SimLex-968, where the twenty entries that contained pairs of identical word, or pairs that were not stemmed properly were manually pruned out.

4.2) GloVe Parsing

To analyze the success of our embeddings for further research, we first had to create an embedding set from our raw corpus using the GloVe software. As mentioned in the Resources section, we acquired three sets of embeddings for our preliminary experimentation: one taken from Facebook which was provided to us by SZTAKI , and two more derived from our Spanish Billion Word (SBW) corpus. The first of these embeddings was provided to us in the form of a text file containing vectors computed with a word2vec algorithm, but the second set we created ourselves by stemming our corpus with a Python script, and passing the word stems through GloVe.

Stemming a language is the process by which verbs and inflected words are reduced to their root, or stem. Examples of this concept in the context of the Spanish language can be seen in Table 3 below. In the examples, it is apparent that stems of the same verb will all be identical, and that accents and inflections will be removed within stems.

English Translation	**Spanish Word**	**Spanish Stem**
to talk	hablar	habl
we talk	hablamos	habl
zoo	zoológico	zoolog
quickly	rápidamente	rapid

Table 3: Stemming examples in Spanish, translated to English

The script that we wrote is designed to read in a text file, iterate through each line in the file using a while loop, and transform each line into a string of stemmed words, which is then written to an output file. Since the corpus given to us was split into one hundred smaller files, we found it easier to modify our original design at the file opening stage. In the new version, a while loop surrounds all of the code from the opening of the corpus file to the writing of the output file, so that the function can iterate over each of the smaller files, and concatenate them into one larger file at the end of the entire process . The resulting seven gigabyte text file contained a fully stemmed corpus and was ready to be passed into GloVe.

4.3) Embedding Evaluation

After processing our stemmed corpus through GloVe, we were left with three full embedding sets which were ready for performance evaluation. In order to test the accuracy of the given vector sets, we used the SimLex-999 data set [19] as our basis for evaluation. Each of the one thousand SimLex-999 pairs was read in, along with one embedding file. We then looped through the SimLex-999 pairs and located the corresponding vectors for each word, if they existed, and computed the cosine similarity between the two. Each word in the vector set was represented as a three hundred dimension vector based on its appearance with other words in a corpus when the vectors were created. A simplified three dimensional representation of two words can be seen in Figure 3.

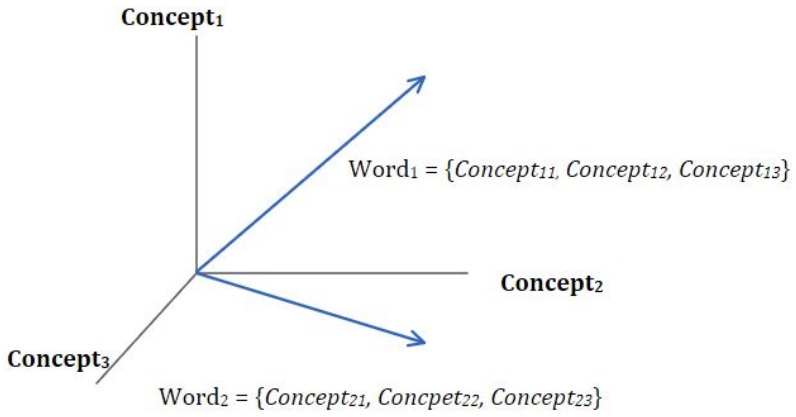

Figure 3: Simplified representation of two words in a 3 dimensional space. Our vector sets were 300 dimensional

The cosine similarity, as shown in Figure 4, produces the similarity between the two sentence vectors.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 4: Cosine similarity formula, whereby A and B would be the vector representations of Word1 and Word2

If one or both words did not have a corresponding embedding, a cosine similarity of 0 was assigned. After computing the similarities for all one thousand pairs, we calculated the Spearman rank-order correlation between the SimLex-999 similarity scores and the computed cosine similarity. The Spearman correlation [24] is the non-parametric version of the Pearson product-moment correlation [25]. Throughout our research, we make note of both Spearman Rho values and p-values. For clarity, a Spearman Rho value is the rank coefficient, or the raw correlation value that denotes the strength and direction of the linear relationship between two variables [24], whereas the p-value is the probability that

of unrelated variables producing the same correlation. The equation for calculating the Rho value is summarized in Figure 5.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Figure 5: Spearman Rho equation for data without tied ranks where d_i = difference in paired ranks and n = number of cases [24]

We evaluated each embedding set two different ways, using partial and exact matching in order to examine potential variances between the stemmed GloVe set, which would almost never have exact matched in the SimLex-999 set due to the nature of its formation, and the other two embeddings. The vector set with the Spearman correlation rho score closest to 1 was the best performing set.

4.4) STS Framework Modifications

Once we selected an embedding set, we were equipped with all of the resources necessary to begin our modifications on the *align-and-penalize* approach framework discussed in Section 3.4. It was our goal to modify the dependencies of the English based framework in order to work with the Spanish specific resources outlined in Section 2.

Additionally, we were notified that the framework we were provided analyses the sentence pairs and outputs a confidence interval between zero and one that measures the similarity between the two sentences. In other words, the sentences within a pair with a confidence value of 0.8 are far more similar to the sentences within a pair with a 0.4

confidence value. After obtaining the set of confidence values for both sets of provided testing data, one from Spanish Wikipedia articles, and the other from news articles, we calculated a Spearman correlation between these values and the gold standard SemEval scores, with the hopes of seeing a high rho value. For comparison, the 2015 winner had a rho value as high as .71 for the Wikipedia dataset and .68 for the news data set [3].

After making necessary modifications, we were then able to begin testing different hyperparameter configurations and embedding sets for optimal STS scoring. In the first set of iterations, we tested both the Wikipedia and the news data on the SBW embedding set with six different combinations of modes (Average, Min, and Max), and the omission of stopwords (True/False). Stopwords are the set of words that do not enhance the meaning of the sentence. English examples would be “I”, “my”, “additionally”, and so on. We found the top three performing configurations within these six, or the three configurations with the highest Spearman rho values, and tested those three configurations with the Facebook embedding set. We were unable to test the GloVe set due to time constraints and file parsing errors due to inconsistencies in word2vec files versus what GloVe outputs.

5.0 – Results and Evaluations

The results from our investigation came in two stages. First, we examined the three sets of embeddings to determine which set performed optimally, and therefore should be used in our research. The second set of investigations came from testing our semantic textual similarity system against the gold standard testing results provided for the 2014 SemEval challenge, and observing the correlation between the two sets. Overall, we found that the Spanish Billion Words embedding set performed optimally but decided to test all three sets in our implementation. When we utilized the SBW embedding set in our similarity system, we saw a very high correlation between the resulting data set and the gold standard.

5.1) Embedding selection

After obtaining all of the embeddings, each set of words and vectors was passed through our evaluation function. The cosine similarity values were compared to the Spanish SimLex-999 gold standard, and a Spearman correlation was calculated. We were able to extract the Spearman p-values and rho values from each correlation to examine the strength, bearing in mind that any value greater than $p = 0.05$ is not considered statistically significant. The results from this experimentation are outlined in Table 4 below.

	Spanish-bwc	*Facebook*	*GloVe*
Exact match // Full SimLex Rho value	0.0624	0.0576	0.0548
Exact match // Full SimLex P-value	0.103	0.061	0.0335
Exact match // Stemmed Simlex Rho value	0.0624	0.0577	0.0548
Exact match // Stemmed Simlex P-value	0.0487	0.0685	0.0835
Partial Match // Full SimLex Rho value	0.0933	0.0571	0.094
Partial Match // Full SimLex P-value	0.0032	0.0713	0.002
Partial Match // Stemmed SimLex Rho value	0.0659	0.095	0.0761
Partial Match // Stemmed SimLex P-value	0.0372	0.0025	0.0162

Table 4: The results of our embedding set performance, analyzed using a Spearman correlation. The top rho value for each set is highlighted in green.

After verifying that all three embedding sets had acceptable Spearman rho values, we were able to conclude that all three embedding sets were adequate for implementation in the rest of our research.

We found that the Spanish Billion Word corpus passed through the word2vec algorithm, as well as the Facebook embeddings performed most optimally on an unstemmed SimLex-999 data set, whereas the GloVe embedding set performed most optimally on the stemmed data. This was expected, given that the embedding set for GloVe is comprised of a stemmed vocabulary, whereas the other two embedding sets were comprised of full words.

5.2) Modified Framework Performance Assessment

Next, we were able to begin running the modified framework with an initial set of hyperparameters, as discussed in section 4.4. We ran all possible hyperparameter combinations with the Spanish Billion Words Corpus on both the news and Wikipedia test data provided to SemEval 2014 challenge participants [1]. The Spearman calculations are summarized in Figure 6 and Table 5 below.

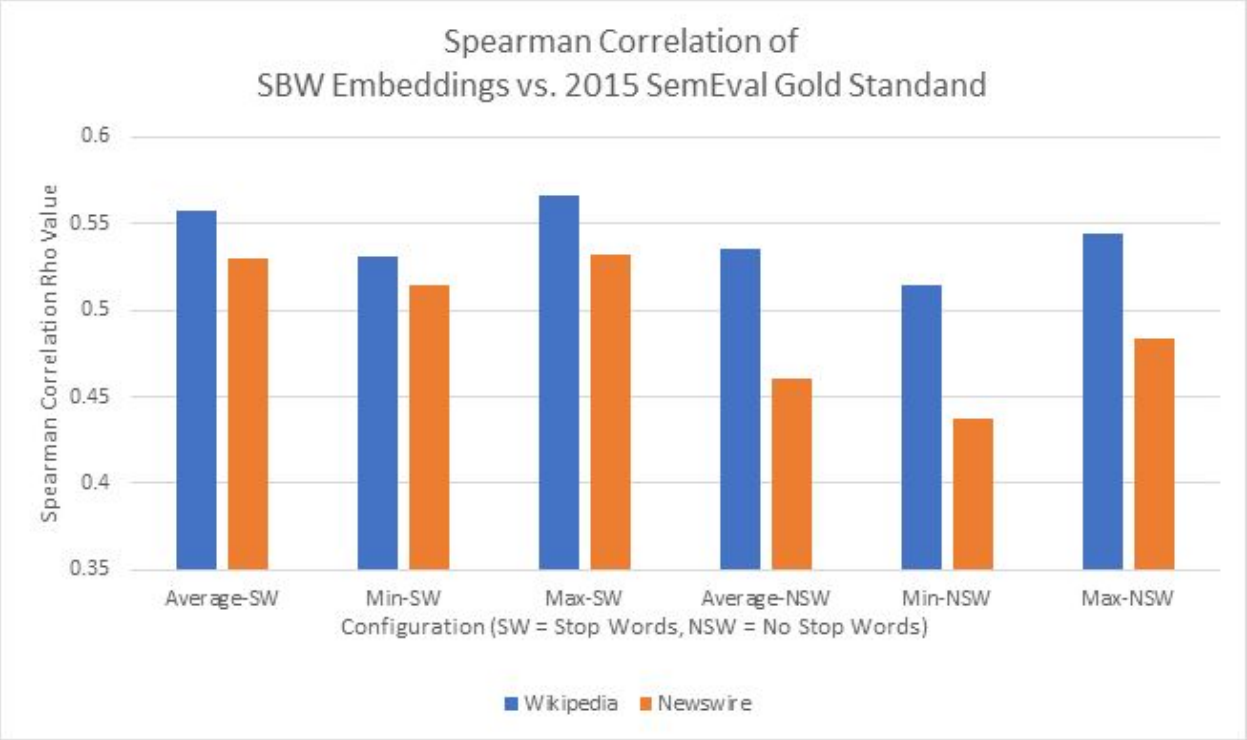

Figure 6: Spearman correlation between SBW embedding set and the Gold Standard used by the 2015 challenge. Data in blue represents the Wikipedia data set. Data in orange represents the newswire data set.

Mode	Stop words	Wikipedia Rho Value	Wikipedia P-Value	Newswire Rho Value	Newswire P-Value
Average	True	0.5527	1.78E-21	0.5299	1.50E-37
Average	False	0.5349	5.57E-20	.05349	5.57E-20
Min	True	0.5304	1.31E-19	0.5149	3.36E-35
Min	False	0.5139	2.57E-18	0.5139	2.57E-18
Max	True	0.5658	1.20E-22	0.5658	1.20E-22
Max	False	0.5442	9.37E-21	0.5442	9.37E-21

Table 5: Spanish Billion Word Corpus Performance using modified hyperparameters with top five rho values highlighted in green

Once all fourteen tests were complete, we calculated the Spearman correlation rho values and p-values and found that the three hyperparameter sets with the highest rho values were created using the (Average, True), (Max, True), and (Max, False) combinations. We then tested the Facebook embeddings using these top three combinations, which are summarized in Figure 7 and Table 6 below.

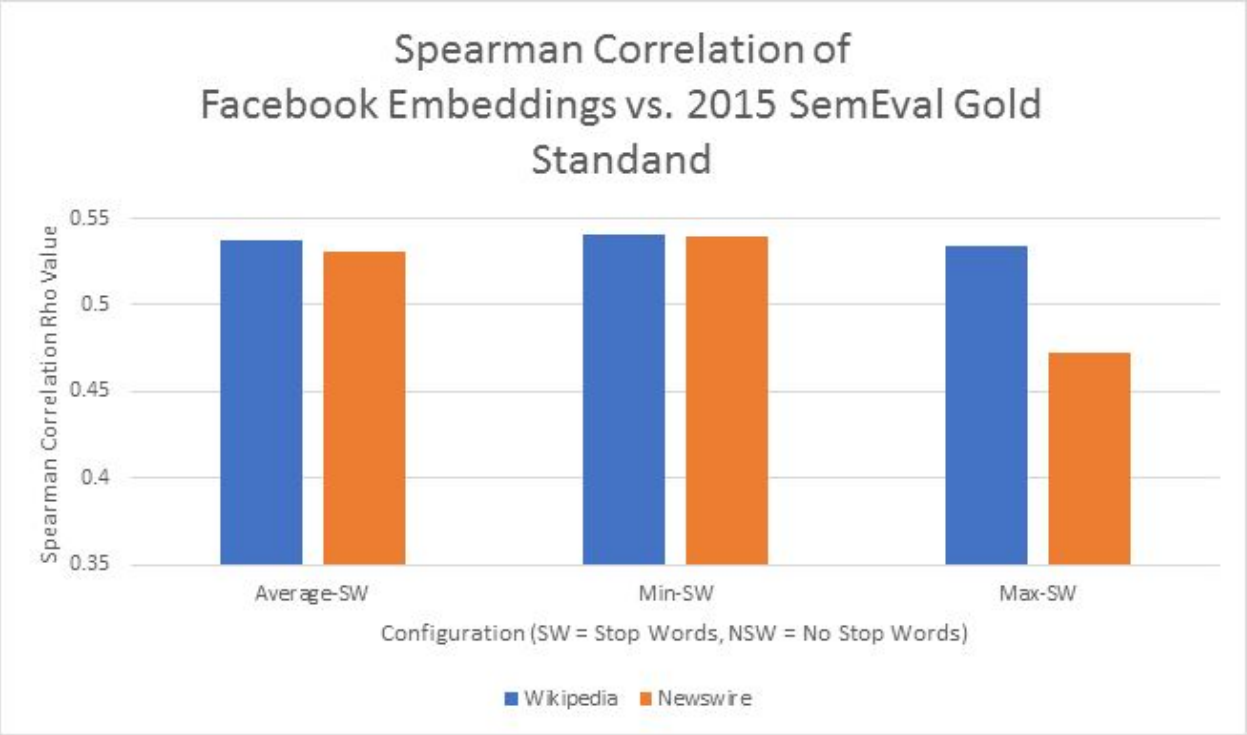

Figure 7: Spearman correlation between Facebook embedding set and the 2015 Gold Standard. Data in blue represents the Wikipedia data set. Data in orange represents the newswire data set.

Mode	Stop words	Wikipedia Rho Value	Wikipedia P-Value	Newswire Rho Value	Newswire P-Value
Average	True	0.5378	3.25E-20	0.5310	1.01E-37
Max	True	0.5402	2.06E-20	0.5402	3.20E-39
Max	False	0.5377	6.98E-20	0.4725	3.61E-29

Table 6: Facebook embedding performance using the top three SBW performing hyperparameter configurations

It is clear from the rho and p-values presented above that the Facebook embeddings performed very similarly to the Spanish Billion Word embeddings. This observation is to be expected given that the initial evaluation of the embeddings revealed similar performance between the two as well.

It was our intention to evaluate these same top three configurations on the GloVe generated embedding set, but due to inconsistencies in the formating of the embedding file, the system was not able to accurately parse and evaluate the data. Additionally, more combinations, including use of the penalty flags, would have been tested but we were limited by time constraints and were unable to do so during the course of the project. We are very happy with the statistical significance of our findings and would have placed around 15th out of the 22 teams which submitted their results during the 2014 challenge.

6.0 – Conclusions & Future Work

This project used our testing and framework to evaluate the Semantic Textual Similarity of Spanish sentences using the structure outlined by the SemEval tasks [2]. Our time and effort produced a new set of SimLex-999 Spanish data, embedding evaluations for the three embeddings available to us, and a functioning and competitive STS evaluation framework. We are proud of our contributions to the progression of Spanish STS, as well as leaving our sponsors with code and resources to continue our research.

6.1) Conclusions

Our work is a valuable contribution, not only to MTA SZTAKI, but to the NLP field as a whole since our adapted SimLex-999 dataset can be used and improved upon for use as a gold standard in Semantic Textual Similarity tasks for Spanish in the future. Additionally, the Human Language Team at MTA SZTAKI has decided to follow suit of our Spanish SimLex-999 and will enlist the help of all members within their research team to create a Hungarian SimLex-999 dataset in the coming months.

The testing and validation of our three embedding sets (Spanish Billion Word, Facebook, and GloVe) proved that they were all significant enough to be used for the purposes of evaluating Semantic Textual Similarity. Finding that all sets are significant not only proves that each set is large enough to cover the SimLex-999 data set adequately, but also indicates that the vectors created using GloVe are accurate enough to represent the

Spanish language as well as vectors Facebook fastText vectors created from Wikipedia, which have been used and verified in other STS tasks.

Lastly, our adaptation of the MathLingBudapest's modular Han et al. framework was a successful competitor against the SemEval participants. With the testing we were able to complete, we know the basic configurations generate sentence scores that are statistically significant to the gold standards made available by the SemEval 2014 Task. We also know that the code could easily be picked up for replication, modification, and future work due to its modular implementation and the documentation we added.

6.2) Future Work

We would like for both the SimLex-999 data set and the modified Han et al. framework be continued and improved upon, either as future undergraduate projects or tasks of the NLP field. We feel that both elements have been adequately established through our time and research and the applications for both have been presented throughout this paper.

For the improvement and enrichment of our SimLex-999 data set, we would like it to be vetted and assessed using the criteria outlined by Leviant and Reichart [26] which outlines a process similar to SimLex-999 [19] including human validation of similarity scores through systems such as Amazon Mechanical Turk. Our ultimate goal would be a Spanish SimLex-999 file with all of the data that the English set has, as outlined in section 3.3, Table 2. Additionally, we believe it would be beneficial to remove the identical pairs as

mentioned in Section 4.1 and instead replacing them with other semantically equivalent words.

In order to enhance the performance of our MathLingBudapest framework, we would like to improve the supporting resources used, increase test coverage, and implement machine learning. Currently, we implement TreeTagger which comes pretrained on a Spanish corpus which we do not have but we would like to train our own POS tagger that way quality can be monitored and tuned.

If future research enables, we would also recommend the usage of a larger corpus, both for the GloVe embedding creation and the POS tagger training. Although 1.5 billion words was big enough for our initial attempt, something on the scale of 3 billion words would be more suitable for optimal results. The last resource modification we would recommend would be a stemmer other than NLTK SnowBall with better performance. We were unable to do so due to financial and time limitations but feel this could have provided an even more competitive embedding set. After making suggested modifications, future researchers would be able to begin to implement the machine learning component of the SZTAKI Han et al. implementation that we were unable to complete during our project.

After making possible resource modifications, we would like to see further testing done with the available hyperparameters. Currently, there are five different binary settings available in the SZTAKI implementation configuration files, making for 32 different test cases. We would want to test all of the combinations on the three different embeddings,

thus increasing that number to 96 different tests in order to find the optimal hyperparameter and embedding combination.

Bibliography

[1] ACL-SIGLEX, "SemEval-2015 Task 2: Semantic Textual Similarity," *SemEval-2015 Task 2*, 2015. [Online]. Available: http://alt.qcri.org/semeval2015/task2/index.php.

[2] ACL-SIGLEX, "SemEval-2014 Task 10," *SemEval-2014 Task 10*, 2014. [Online]. Available: http://alt.qcri.org/semeval2014/task10/index.php?id=sts-es.

[3] ACL-SIGLEX, "SemEval-2012: Semantic Evaluation Exercises," *SemEval-2012*, 2012. [Online]. Available: https://www.cs.york.ac.uk/semeval-2012/.

[4] ACL-SIGLEX, "SemEval-2016: Semantic Evaluation Tasks," *SemEval-2016*, 2016. [Online]. Available: http://alt.qcri.org/semeval2016/index.php?id=tasks.

[5] J. Acs and A. Informatics, "MathLingBudapest: Concept Networks for Semantic Similarity," no. SemEval, pp. 138–142, 2015.

[6] E. Agirre, "SemEval-2015 Task 2: Semantic Textual Similarity , English , Spanish and Pilot on Interpretability," *SemEval2015*, no. SemEval, pp. 252–263, 2015.

[7] E. Agirre *et al.*, "SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation," *Proc. 10th Int. Work. Semant. Eval.*, pp. 497–511, 2016.

[8] E. Agirre, D. Cer, M. Diab, and B. Dolan, "SemEval-2015 Task 2," 2015.

[9] M. Alan, "Turing. Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[10] J. Allen, *Natural Language Understanding*, 2nd ed. Redwood City, CA: Benjamin-Cummins Publishing Co. Inc, 1995.

[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *arXiv:1607.04606v1 [cs.CL]*, 2016.

[12] U. of Cambridge, "SimLex-999," 2016. .

[13] C. Cardellino, "Spanish Billion Words Corpus and Embeddings." p. Spanish Billion Words Corpus and Embeddings, 2016.

[14] D. Chen and M. Christopher, "A Fast and Accurate Dependency Parser using Neural Networks.," *A Fast and Accurate Dependency Parser using Neural Networks.*, 2014. .

[15] J. Cheng, Z. Wang, J.-R. Wen, J. Yan, and Z. Chen, "Contextual Text Understanding in Distributional Semantic Space," *Proc. 24th ACM Int. Conf. Inf. Knowl. Manag. - CIKM '15*, pp. 133–142, 2015.

[16] Google, "Vector Representations of Words," *TensorFlow*, 2016. [Online]. Available: https://www.tensorflow.org/tutorials/word2vec#motivation_why_learn_word_embeddings.

[17] L. Han, A. Kashyap, T. Finin, J. Mayfield, and J. Weese, "UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems," *Proc. 2nd Jt. Conf. Lex. Comput. Semant.*, vol. 1, pp. 44–52, 2013.

[18] P. Hayes and J. Carbonell, *A Tutorial on Techniques and Applications for Natural Language Processing*. 1983.

[19] F. Hill, R. Reichart, and A. Korhonen, "SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation," 2014. [Online]. Available: https://github.com/se4u/mvlsa/blob/master/res/word_sim/SimLex-999.README.

[20] IBM, "What is big data?," *Bringing big data to the enterprise*, 2016. .

[21] S. Jimenez *et al.*, "UNAL-NLP: Combining Soft Cardinality Features for Semantic Textual Similarity, Relatedness and Entailment," *Proc. 8th Int. Work. Semant. Eval. (SemEval 2014)*, no. 1, pp. 732–742, 2014.

[22] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," 2016. [Online]. Available: https://github.com/facebookresearch/fastText.

[23] F. B. Kessler, "About MultiWordNet," *MultiWordNet*, 2008. [Online]. Available: http://multiwordnet.fbk.eu/english/home.php.

[24] Laerd Statistics, "Spearman's Rank-Order Correlation," 2016. [Online]. Available: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php.

[25] Laerd Statistics, "Pearson Product-Moment Correlation," 2016. [Online]. Available: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php.

[26] I. Leviant and R. Reichart, "Judgment Language Matters: Towards Judgment Language Informed Vector Space Modeling," 2015.

[27] Y. Li, D. Mclean, Z. Bandar, J. D. O. Shea, and K. Crockett, “Sentence Similarity Based on Semantic Nets and Corpus Statistics,” pp. 1–35.

[28] C. D. Manning and A. Gravano, “Turn-taking and affirmative cue words in task-oriented dialogue,” *Diss. Abstr. Int. B Sci. Eng.*, vol. 70, no. 8, p. 4943, 2010.

[29] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. Cambridge: Massachusetts Institute of Technology, 1999.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” *Nips*, pp. 1–9, 2013.

[31] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, “Natural language processing: an introduction.,” *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 544–51, 2012.

[32] P. Nakov, “Siglex - ACL Special Interest Group,” 2016. [Online]. Available: http://siglex.org/.

[33] N. Nesselhauf, “Corpus Linguistics: A Practical Introduction,” vol. 2005, no. October 2005, p. 32, 2011.

[34] M. Padró, M. Ballesteros, H. Martínez, and B. Bohnet, “Finding Dependency Parsing Limits over a Large Spanish Corpus,” *Proc. Sixth Int. Jt. Conf. Nat. Lang. Process.*, no. October, pp. 942–946, 2013.

[35] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation.”

[36] H. Schmid, “TreeTagger -- a part-of-speech tagger for many languages,” *2015*. [Online]. Available: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

[37] L. Tan, C. Scarton, L. Specia, and J. van Genabith, “SAARSHEFF at SemEval-2016 Task 1: Semantic Textual Similarity with Machine Translation Evaluation Metrics and (eXtreme) Boosted Tree Ensembles,” *Proc. 10th Int. Work. Semant. Eval.*, pp. 628–633, 2016.

[38] Princeton University, “About WordNet,” *WordNet*, 2010. [Online]. Available: http://wordnet.princeton.edu.

[39] W. Xu, C. Callison-Burch, and B. Dolan, “SemEval-2015 Task 1,” 2015. [Online]. Available: http://alt.qcri.org/semeval2015/task1/.