A Method for the Evaluation of the Semantic vs Syntactic Localization of Sentence Embeddings

Lyndon White

Friday 24th July, 2015

.

1 Introduction

Here we will define a sentence embedding as the embedding of a sentence. This is often called a phrase in the work of Sorcher et Al[1, 2, 3, 4, 5], and is one form of what is called a paragraph in [6]. A sentence is a single idea, thus is a valid target for TRAVEL TO THE MOON.

Sentences are said to be semantically equivalent if they each imply the other – the relationship is that of bidirectional entailment.

The paraphrases from the MSRPC were judged by the human raters to have the same high-level meaning, and thus show "mostly bidirectional entailment" [7]. That is to say, that while each sentence may contain information which is not implied by the other, the core meaning of the sentences is entailed by both. [7] provides the examples of:

> Charles O. Prince, 53, was named as Mr. Weill's successor.
>
> Mr. Weill's longtime confidant, Charles O. Prince, 53, was named as his successor.

While additional information is present, each sentence implies most of the meaning of the other. Thus while not semantically equivalent, they are semantically close.

It should be noted that that semantic similarity is often defined differently for sentences as for words. While semantic similarity for sentences is defined in terms of shared meaning and mutual entailment, semantic similarity for words can be defined in terms of shared properties[8]. For example: "rise" and "fall" are antonyms, but are under the aforementioned definition for word semantic similarity are very similar: they both describe an vertical change in (potentially metaphorical) position. However, the sentences "The share price is predicted to rise." and "The share price is predicted to fall." are not semantically similar sentences, as they do not imply each other - in-fact each implies that the other is false. This sentence definition of semantic equivalence can be seen to be essential in applications such as machine translation.

2 Background and Motivation

2.1 Semantic Tasks

Many of the key tasks in natural language processing and generation have strong semantic components:

- Translation
- Sentiment Recognition
- Indexing and Retrieval
- Summerisation
- Automatic Paraphasing[9]

2.2 Embeddings

Various sentence embeddings are often called Semantic Vector Space Representations, such as in [9]. This paper aims to provide a method to assess the accuracy of that statement. Such a tool will allow us to assert whether the vector representation of a given method is more influenced by the structure (syntax) of the sentence embedded, or of its meaning (semantics).

2.3 Motivating better use of Semantic Resources in embedding creation

[10]

2.4 Existing Corpora

2.4.1 Paraphrase Corpora

P4P, MSRP

2.4.2 Finding Structural Matches in Large Corpora

Inspection of large corpora will reveal that structurally identical sentences are rare. While Large corpora contain many thousands of sentences, very few have the same structure, or even same order of POS tags. For example, MASC[11] (Manually Annotated Subcorpus of the Open American National Corpus) contains 34,535 sentences, only 4,230 share a set and ordering of POS tags with another. Further: if trivial sentences with 5 or fewer tokens are excluded there are just 634. If sentences where 30% or more of the words are proper nouns, dates, times, or numbers are also excluded (thus removing sentences such as "'Posted by : tomemos | Sunday , 09 May 2010 at 09:16 AM'), only 316 sentences remain. Only a subset of these will be structurally identical when considered under a parse tree.

Thus it is not surprising that MSRPC, which only contains 7800 sentences, contains 5 structurally identical sentences in two structural classes, other than paraphrases or differing only numerically. Even these sentences are highly similar in topic and word usage. Class 1: {’Schering-Plough shares fell 72 cents to close at $ 18.34 on the New York Stock Exchange .’, ’Wal-Mart shares rose 16 cents to close at $ 58.28 on the New York Stock Exchange .’, ’Xerox shares rose 2 cents to close at $ 11.45 on the New York Stock Exchange .’}, Class 2: {’Jason Giambi capitalized with an RBI single to center .’, ’Randall Simon followed with an RBI single to right .’}.

Similarly, there is only one structural match between MSRPC and the MASC corpus: MSRPC contains: “I’m never going to forget this day.” which is a structural match for the sentence: “I’m not going to allow that question.” from MASC. Thus it is not viable to find naturally occurring sentences of the same structure, but different meaning.

2.4.3 Lexical Substitution Corpora

.

CoInCo (Concepts in Context) is a manually created all-words lexical substitution corpus[12]. Approximately 2500 sentences from a subset of MASC (Manually Annotated Subcorpus of the Open American National Corpus), were given to human annotator, who were tasked with listing all the word synonyms that could replace words. CoInCo only handles synonym substitutions, though an effort could be made to automatically extend it to antonyms (and other lexical relations), by mapping substitution lemmas from CoInCo on to WordNet synsets. According to [12] CoInCo is the only large scale, manually created corpus of its type, though other single word substitution corpora exist.

2.4.4 What is required

What is required is a corpus containing

3 Creation of Corpus

3.1 Base Corpus

The Microsoft Research Paraphrase Corpus is used as a base source of sentences, and a ground truth for their semantic equality. The sentences with paraphrases provided are combined from the training and testing sets. This gives 7,800 sentences for evaluation. Of these 311 are repeated – that is they have multiple different paraphrases specified – no special handling is done for these cases. The sentences are modified by the procedure described below to create semantically equivalent and semantically distinct versions.

3.2 Tokenization

The first step is to tokenize the sentence. This was achieved using the NLTK[13] Treebank Word Tokenizer. This tokenizer is based on regular expressions. It splits the sentences in to words, punctuation elements, and also separates contractions: "don't" becomes "do n't". Tokenization is a fairly simple task, accomplished by these regular expressions.

3.3 POS Tagging and Restriction of Auxiliary Verbs

The second step is to tag the the word tokens with their parts of speech (POS). The Stanford POS Tagger[14] was used via NLTK[13] interface, to accomplish this. The Stanford POS Tagger, has a 97.24% accuracy on the Penn Treebank Wall Street Journal test set[14]. It is one of the best POS taggers available.

Though it does still make some mistakes, for example, in the sentence: "Cadets were ticketted for drinking alcohol." drinking is mistaken for a noun, when it should be a verb.

The sentences are tagged with the Penn Treebank tagset[15]. This tagset contains 37 POS tags. Of interest to this work are the tags for nouns and verbs. Valid noun tags for transformation are NN and NNS, which covers singular, mass and plural nouns. The proper noun tags NNP and NNPS are not valid for transformation and are excluded. Valid verbs are marked those marked with all verb tags (VB,VBD,VBG, VBN, VBZ, VBP), except for models (MD) and other auxiliary verbs.

Model verbs and auxiliary verbs (including Models) are normally inverted by inserting a not after the verb, or equivalently a contraction n't.[16] This is forbidden by the guideline of not changing the structure in the generated sentences. Thus model verbs and auxiliaries are excluded.

While the POS tagger captures models with the MD tag, the other auxiliary verbs are not caught. [15] states they are to be handles as other verbs. Some of them also have non-auxiliary senses, for example had in "he has gone" is an auxiliary[16], but in "He has a dog" it is not. WordNet also does not differentiate auxiliaries from other verbs, and so in both cases suggests that antonym for "has" is "lacks". To avoid any confusion of this sort auxiliary verbs are blocked using a blacklist. This blocks changes to: "be", "am", "are", "is", "was", "were", "being", "can", "could", "do", "did", "does", "doing", "have", "had", "has", "having", "may", "might", "must", "shall", "should", "will", and "would".

WordNet uses must simpler parts of speech tags, as it only considered lemmas. WordNet POS tags are: noun, verb, adverb and adjective. The Penn Treebank POS tags can be simplified down to them. Further more, the additional information captured in the Penn Treebank Tags, is sufficient to allow recover the full form an a lemma generated from WordNet.

3.8).

3.4 WordNet Lookup

Synsets are looked up by lemma name from Wordnet.

The lemma name is generated using Morphy, thought the NLTK interface.

3.5 Phrase Detection

Certain sets of words are best treated as a single unit, this paper will call these phrases. WordNet Version 3.0 contains 64,331 such phrases.

Consider the word sequences: “chief financial officer”, and “police officer”.

A synset exist containing “chief_finial_officer” and “CFO”, another exist containing: “police_officer”, “officer” and “policeman”. If phrases were handled as word sequences “chief financial officer” could have the officer replaced, to get: “chief financial policeman”, or even: “police officer” could become “police police officer”.

If phrases were handled as words, “policeman” could become “police officer”. This adds a word, violating the constraint of not changing the sentence’s structure. To avoid all these issues entirely, we forbid the substituting for any words in a phrase, as well as forbidding substituting a phrase for any words.

Continuous dictionary phrases are detected using a sliding window of width 3 and 2 words across the sentences. The words in the window are then checked to see if they form a phrase known to WordNet. If they do then they are blocked from substitution. This blocks all continuous dictionary phrases.

While this blocks all dictionary phrases it does not handle other kinds. Several other kinds of phrases have been distinguished as having distinct meaning. Such as the skip-bigrams considered by [17]. The necessity of avoiding substitutions with these is less clear. As they are not considered as clear single lexical entity (unlike dictionary phrases), they may be sufficiently handled by word sense disambiguation.

3.6 Word Sense Disambiguation

Word sense disambiguation is used to select the correct sense of the word being substituted for – so that it synonym or antonym is the same sense. This is used for example to ensure that the synonym of “bank”, as in a financial institution, is not “shore” as in the edge of a body of water[8]. While several methods were considered for performing the word sense disambiguation, it was resolve to use the simple Most Frequent Sense (MFS).

The MFS is a naïve method for determining word sense. It functions by always choosing the word sense that occurs most often – without regard for context, beyond the POS tag of the word. This method is almost certain to make some mistakes, however more sophisticated algorithms have been shown to offer little improvement over it.

BabelNet[18] is a multilingual extension of WordNet that has seen significant use as a target for the evaluation of word sense disambiguation methods. In the

English BabelNet WSD subtask at [19], the MFS obtained a F1 score of 66.5%, the best competing algorithm scores 68.6%. Since then, new methods have improved that to [20] 71.5% – which exceeds MFS by 5%. On a similar task at [21], no entry exceeded the MFS F1 score of 67.5%. As simple MFS remains one of the most competitive methods it is used in this system.

3.7 Substitution Generation

3.7.1 Noun Synonyms

As discussed above, all words sorted in WordNet are stored in synsets[22]. A synset contains many lemmas all with the same semantic meaning. Any given word may have many word-senses, each word-sense for that word belongs to a different synset. All the lemmas within a synset are synonymous. As we are ignoring word-sense disambiguation, the possible synonyms for a word are the union of all lemmas of all synsets with the correct POS tag – in this case noun.

3.7.2 Verb Antonyms

3.8 Un-lemmatizing

As the words generated from WordNet are lemmatized, this process must be reversed to restore them to their grammatically correct from for the context. For example, the verbs, “rise”, “rose”, and “rising” are all mapped to the lemma “rise”, from this lemma we generate the antonym: “fall”, to put it into the context of the first word, it needs to be mapped to “fall”, “fell”, and “falling” respectively. Similarly, for nouns lemmatization removes plurals. The needed information for both is captured in the parts of speech tags.

The Penn Treebank POS tags captures the information required to go from a lemma to the correct form. This can be used with the heuristic conjugation and pluralization methods from the Pattern.en library [23] to correct the generated lemma to the appropriate form. The mappings used are shown in figure 1. It can be noted that no processing is done on the Verb, non-3rd person singular present case (VBP), this is because with the exception of the various forms of the verb “be”, which are excluded earlier as and auxiliary/model, the non-3rd person singular present is always the same as the base form of the verb[24, p. 84]. These rules have been found to be generally sufficient to ensure grammatical text.

To ensure against any failure in the heuristic rules based un-lemmatization the validity of the generated and un-lemmatized word is checked against the “british-english-insane” collection of words from the Spell Checker Orientated Word Lists (SCOWL)[25]. Any generated words which fail this test are discarded.

During this step, the initial letter capitalization of any generated word is matched to that of the base word. While no method evaluated below makes

Sentence	base word	generated word	POS	POS meaning[15]	Pattern.en method to un-lemmatize
The share price will rise.	**rise**	**fall**	VB	Verb, base form	–
The share price rose.	rose	fell	VBD	Verb, past tense	`conjugate(lemma, tense=PAST)`
The share price is rising.	rising	falling	VBG	Verb, gerund or present participle	`conjugate(lemma, tense=PRESENT, aspect=PROGRESSIVE)`
The share price normally rises.	rises	falls	VBZ	Verb, present tense, 3rd person singular	`conjugate(lemma, tense=PRESENT)`
The share price has risen.	risen	fallen	VBN	Verb, past participle	`conjugate(lemma, tense=PAST, aspect=PROGRESSIVE)`
The share prices rise.	**rise**	**fall**	VBP	Verb, non-3rd person singular present	–
The car is hot.	**car**	**automobile**	NN	Noun, singular or mass	–
The cars are hot.	cars	automobiles	NNS	Noun, plural	`pluralize(lemma)`

Figure 1: POS Tags for various forms of words. The WordNet lemma is in **bold**.

use of capitals they are preserved for ease of future comparisons with methods which do.

[1]

3.9 Indefinite Article Correction

A word substituted for may have been preceeded by an indefinate article. Depending on the vowel sound at the start of the next word, it may no longer be appropriate[24, p. 1618]. This case is detected and rectified, again making use of the Pattern.en library[23]. This correctly handles phonetic cases such as “an honest man” and “a unit of measure”. Case is again preserved. This is the final step in the generation of semantically altered sentences.

[1] Pattern.en Verb Conjugation methods use are described in detail in the documentation http://www.clips.ua.ac.be/pages/pattern-en#conjugation

4 Examples of Constructed Evaluation Sentences

Change	**Sentence**	**Wiki PV-DM Concatenting Model Distance**
Base Sentence	However , other unions including the powerful CGT remained opposed to the reform and demanded the government begin fresh negotiations with them .	0.00
Paraphrase	The powerful CGT and other unions remained opposed to the plans , however , and demanded the government renegotiate the reform with them .	**0.35**
1 Noun Synonym	However , other **brotherhoods** including the powerful CGT remained opposed to the reform and demanded the government begin fresh negotiations with them .	*0.27*
2 Noun Synonym	However , other brotherhoods including the powerful CGT remained opposed to the reform and demanded the **authorities** begin fresh negotiations with them .	*0.25*
3 Noun Synonym	However , other brotherhoods including the powerful CGT remained opposed to the reform and demanded the authorities begin fresh **dialogues** with them .	0.43
1 Verb Antonym	However , other unions including the powerful CGT remained opposed to the reform and **obviated** the government begin fresh negotiations with them .	*0.36*
2 Verb Antonym	However , other unions including the powerful CGT remained opposed to the reform and obviated the government **end** fresh negotiations with them .	*0.36*
3 Verb Antonym	However , other unions including the powerful CGT **changed** opposed to the reform and obviated the government end fresh negotiations with them .	*0.43*
4 Verb Antonym	However , other unions **excluding** the powerful CGT changed opposed to the reform and obviated the government end fresh negotiations with them .	*0.46*

5 Evaluation

A given model may be evaluated on how semantically close, or distantly it places sentences which are semantically equivalent, or not. As discussed, replacing a noun with its synonym produces a semantically equivalent sentence – both the original sentence, and the modified sentence entail the other. Conversely, replacing a verb with its antonym generally will produce a semantically different sentence, often the the modified sentence will entail the converse of the original and visa-verse. As a baseline for a reasonable distance under a model, the distance to the gold standard, semantically equal, paraphrase is computed.

5.0.1 Scoring

By comparing the distance from the base sentence embedding to the paraphrase and modified sentence embeddings, the success of the modal at semantic localization is evaluated. It is assessed on keeping close semantically similar sentences, and distant semantically distinct sentences. A semantically different sentence must be further away that a semantically equivalent one, thus the sentences modified by replacing verbs with their antonyms must be more distant than the semantically equal paraphrase. Conversely, the semantically identical similar form should be no further away than the paraphrase. By counting the portion of the evaluation corpus is correctly placed under this definition, the models semantic semantic accuracy with respect to fine-grained changed is scored.

5.0.2 Distance

5.1 The Models

For demonstration purposes, several models are evaluated below.

5.1.1 RAE

5.1.2 Doc2Vec: PV-DM concatenating

5.1.3 Doc2Vec: PV-DM concatenating

6 Results and Discussion

7 Conclusion and Future Work

References

[1] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars," in *ACL*, 2013.

[2] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *Advances in Neural Information Processing Systems 24*, 2011.

[3] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, “Parsing natural scenes and natural language with recursive neural networks,” in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 129–136, 2011.

[4] R. Socher, C. D. Manning, and A. Y. Ng, “Learning continuous phrase representations and syntactic parsing with recursive neural networks,” in *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, pp. 1–9, 2010.

[5] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, “Semi-supervised recursive autoencoders for predicting sentiment distributions,” in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[6] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1188–1196, 2014.

[7] W. B. Dolan and C. Brockett, “Automatically constructing a corpus of sentential paraphrases,” in *Third International Workshop on Paraphrasing (IWP2005)*, Asia Federation of Natural Language Processing, 2005.

[8] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[9] M. Iyyer, J. Boyd-Graber, and H. D. III, “Generating sentences from semantic vector space representations,” in *NIPS Workshop on Learning Semantics*, 2014.

[10] M. Yu and M. Dredze, “Improving lexical embeddings with semantic knowledge,” in *Association for Computational Linguistics (ACL)*, pp. 545–550, 2014.

[11] R. J. Passonneau, C. Baker, C. Fellbaum, and N. Ide, “The masc word sense sentence corpus,” in *Proceedings of LREC*, 2012.

[12] G. Kremer, K. Erk, S. Padó, and S. Thater, “What substitutes tell us-analysis of an” all-words” lexical substitution corpus,” in *Proceedings of EACL*, 2014.

[13] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. " O’Reilly Media, Inc.", 2009.

[14] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, “Feature-rich part-of-speech tagging with a cyclic dependency network,” in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 173–180, Association for Computational Linguistics, 2003.

[15] B. Santorini, "Part-of-speech tagging guidelines for the penn treebank project (3rd revision)," 1990.

[16] A. Radford, *English syntax: An introduction.* Cambridge University Press, 2004.

[17] W. Yin and H. Schütze, "An exploration of embeddings for generalized phrases," *ACL 2014*, p. 41, 2014.

[18] R. Navigli and S. P. Ponzetto, "Babelnet: Building a very large multilingual semantic network," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 216–225, Association for Computational Linguistics, 2010.

[19] R. Navigli, D. Jurgens, and D. Vannella, "Semeval-2013 task 12: Multilingual word sense disambiguation," in *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, vol. 2, pp. 222–231, 2013.

[20] P. Basile, A. Caputo, and G. Semeraro, "An enhanced lesk word sense disambiguation algorithm through a distributional semantic model," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, (Dublin, Ireland), pp. 1591–1600, Dublin City University and Association for Computational Linguistics, August 2014.

[21] A. Moro and R. Navigli, "Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking," *Proceedings of SemEval-2015*, 2015.

[22] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[23] T. De Smedt and W. Daelemans, "Pattern for python," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2063–2067, 2012.

[24] R. Huddleston and G. Pullum, *The Cambridge Grammar of the English Language.* Cambridge textbooks in linguistics, Cambridge University Press, 2002.

[25] K. Atkinson, "Scowl (spell checker orientated word lists)," 2011.