Measuring the Limit of Semantic Divergence for English Tweets

Dwijen Rudrapal
CSE Department
NIT Agartala, India
dwijen.rudrapal@gmail.com

Amitava Das
CSE Department
IIIT Sricity, India
amitava.das@iiits.in

Abstract

In human language, an expression could be conveyed in many ways by different people. Even that the same person may express the same sentence quite differently when addressing different audiences, using different modalities, or using different syntactic variations or may use different set of vocabulary. The possibility of such endless surface form of text while the meaning of the text remains almost same, poses many challenges for Natural Language Processing (NLP) systems like question-answering system, machine translation system and text summarization. This research paper is an endeavor to understand the characteristic of such endless semantic divergence. In this research work we develop a corpus of 1525 semantic divergent sentences generated from 200 English seed tweets.

1 Introduction

A sentence could be expressed in innumerable ways without changing its meaning. Different people may express a sentence in different ways and even that the same person may express a sentence quite differently when addressing different audiences, using different modalities, or tackling different tasks (Dewaele, 1999). Possible number of restatement of the meaning of a sentence in a language is huge (Bell, 1995). These variations may be for different writing styles of different people with diverse background and situations (Karlgren, 1996) (Chambers J.K, 2006), number of polysemous words present in the text, longer/deeper syntactic relations for more than one word or presence of more adjective/ adverb than noun and verb in sentence. For example:

Expression 01: After that game I think Harding deserves another one.

Expression 02: Harding deserves another game after that.

Expression 03: I think Harding should get another chance after that game.

Above expressions are restating the same meaning in different ways. These deviations raise a fundamental question that, *are there any set of rules that govern different factors to determine the degree of semantic divergence of an expression?* To find out the answers i.e. measuring the semantic divergence of English sentence, we prepare a corpus of sentence variants by preserving the meaning of a sentence. In this research work, we concentrate our effort on English social media text specifically on tweets as diverse syntactic variations of texts are more prevalent in informal setting and explore different factors other than situational and personality factors which make presentation variations of a tweet.

The structure of this research paper is as follows. First, we draw the problem definition in section 2, discuss related research work in section 3, followed by corpus development process in section 4. In section 5, we present the explored features behind the semantic variations of an English tweet, explain thematic closeness of variations in section 6. We carry out detail analysis on results in section 7. Finally in section 8, we conclude our work with performance evaluation along with noted limitations and mentioning future scope of work.

2 Problem Description

Semantic diversity nature of natural language is highly influential for a broader range of NLP applications (Madnani and Dorr, 2010). In this regard, we discuss some state-of-the-art NLP sys-

Proceedings of Recent Advances in Natural Language Processing, pages 618–624,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_080

tems in this section to draw our research problem.

The performance of one automated NLP system is evaluated by comparing human annotators output against the system generated output. For example, evaluation method for machine translation system, text summarization system, etc. In such evaluations, system outputs are evaluated against reference outputs by measuring the n-gram overlap between them. This measure is completely dependant on exact or partial matching of candidates in both outputs. However, a single reference output may not reflect the best result of matching due to the semantic variation even though both outputs convey the same semantic content. The evaluation process needs to consider all the possible semantically same variants while matching to award better credit. Legitimate variation of a query may retrieve more relevant information. So, the automatic generation of query variants for submission to information retrieval systems should consider all the possible query variants to reach optimal performance.

Our research problem is to explore several important features responsible for arranging information of a tweet into a series of alternatives to reinterpret the tweet in different ways. To address the issue we develop one corpus of semantically same sentences for tweets and present theoretical explanations and evidence in support of the explored features.

3 Related Work

Most of the research work on semantic divergence of expressions focused either on language translation task or to measure formality of document or sub-document units like sentences considering their important applications. But the reinterpretation of a sentence to infer its meaning in different way is also an important research challenge as discussed in section 2. Unfortunately, very few research work have done in this domain. In this section we discuss significant research work on semantic divergence in above two domains.

The work (Hirakawa et al., 1994) proposed an interactive rewriting tool as a part of Japanese-to-English machine translation, where a sentence rewritten based on morphological and syntactical information of the source text. Grammatical transformation of one sentence like syntactic and semantic structure used in the work (Mitamura and Nyberg, 2001) for phrase re-arrangement to reinterpret the sentence by preserving its meaning for translation. Proposed model (Galley and Manning, 2008) (Tomoki Fujita and Nakamura, 2013) also used rewriting tool using phrase reordering model to improve machine translation system for Chinese-English and Arabic-English languages. The work (Wang, 2013) proposed a text rewriting decoder, works on the sentence level features like the language model score of the whole sentence. The work (He et al., 2015) introduced grammatically and meaning-preserving syntactic preservation rules such as verb, noun and clause re-ordering on constituent parse trees for Japanese to English machine translation. Query rewriting process in the work (Riezler and Liu, 2010) generated a set of alternative queries which are semantically the same to achieve the best search result from large amounts of user query logs.

Other than machine translation, semantic divergence has a role in formal or informal writing style. The study (Dewaele, 1999) discussed the degree of formality in linguistic expression based on different situational and personality factors. In the work (Lahiri et al., 2011) authors presented an annotated corpus of 600 sentences with variations in formality on a Likert scale (Likert, 1932) of 1-5. Variants of a sentence with the same meaning is possible with re-arranging the grammatical organization of that sentence was discussed in the article (Rafajlovicova, 2002).

In our current research work, we develop a corpus of semantically same sentences for English tweets and made theoretical study to explore possible features which can be utilized in query variants, enlarge sparse human reference data in evaluation and machine translation evaluation system.

4 Corpus Preparation

To develop a corpus of semantic divergent sentences for tweets, initially we take 200 unique English tweets, randomly chosen from SemEval-2015 Task 1 [1] tweet corpus and from Ritter[2] corpus. These tweets are in raw form and include typos, bad grammar, usage of slang, presence of unwanted content like URLs, emoticons, etc. We extract meaningful text content from each tweet by filtering out html entities like &lt, &gt, &amp, emoticons, embedded URLs, usernames, replaced

[1] http://alt.qcri.org/semeval2015/task1/

[2] www.github.com/aritter/twitter_nlp

#hashtag by hashtag, split multiple attached words like *GoldenGlobes* into *Golden Globes*. Apostrophes and punctuation are kept unchanged to retain tweets meaning intact and resourced to help the annotation task for appropriate semantic variations.

We develop one web-page and upload these preprocessed 200 tweets for collecting possible semantic variations for each tweet without altering meaning. Every tweet is hyper-linked to a page where a user can rewrites the tweet and submit. Every submission gets updated instantly and a future user can sees all the previous re-written expressions for that particular tweet. Thus information redundancy is avoided at the time of annotation. While re-writing, when a user see that all the possible variations for that expression are already listed, the user clicks on a check box labeled as "*No more variation possible*" message and submits. This message signifies the limit of the semantic variations of a particular tweet. To be persistent we record *No more variation possible* at least from two different users until we stop showing the particular tweet to other annotators.

Through this web-page, over the course of about 6 weeks, we collect divergent sentences for tweets by human annotators. Human annotators are Post-Graduate students and native English speaker. Total 21 users participated in the annotation task for given 200 tweets. We have collected 1,525 sentences as variants of semantically same sentences for 200 tweets. The highest number of variations for a tweet found was 24 while the lowest number was 4. We also observe that all the tweets reach the message "no more variation possible" two/multiple times ensuring that all the possible variations of a tweet obtained.

5 Measuring of Semantic Divergence

We analyze the developed corpus to determine the important features, responsible for semantic divergence of an English tweet. We observe that there are some grammatical devices like structural, semantic, pragmatic, and textual factors, used for rearranging the information in the message to express it differently. In our work, we select some features such as synonym, passivization, clause reordering, idioms and phrase, cleft, negation and use of non-impact words and set up one questionnaire to manually annotate every variation of a tweet by tagging with one of the selected features. The questionnaire collects responses from two human annotators for two questions. One, does re-written expression represent the same meaning as in original tweet? Two, which feature/features are use to re-written this expression? The response for the first question is yes when the meaning is fully preserved otherwise no. In response to the 2nd question, annotators select one or multiple features from listed features. We calculated inter-annotator agreement statistics (IAAS) on each feature of annotations. Feature wise distribution of the semantic variations for tweets and IAAS is represented in table 1. Observe that there is high agreement on most of the features due to the straight-forward meaning and use of features in variant expressions. The IAAS for the responses of question no.01 is as high as 99.125, reveals that almost all of variants in the corpus preserved the meaning of original tweet while re-written. Theoretical explanations of each feature with an example is discussed in this section to justify the role of the features.

Features	Instance in corpus (%)	IAAS
Activization/ Passivization	5.32	98.67
Synonyms	34.82	98.89
Idioms & phrase	0.4	98.92
Inter-Clause re-ordering	18.29	92.7
Sentence Type Transformation	12.98	100.0
Cleft	1.25	96.7
Addition/deletion of non-impact words	15.4	96.0
Using Multiple features	11.34	100.0

Table 1: Feature wise distribution of semantic variation

5.1 Activization and Passivization:

The representation of a sentence from active to passive or passive to active voice allows structural re-organization of the expression without any alteration of meaning. Active voice describes a sentence where the subject performs the action stated by the verb. Passive voice describes a sentence where the subject is acted upon by the verb. For example:

Original tweet: My phone NEVER sends me Amber Alerts.
Variant: Amber alert is never received by my phone.

In the above variant example, subject (My Phone) is being acted on by the verb (receive) and comes after the action in the sentence. In our developed corpus a total of 81 instances (5.32% of total variants) observe in this category.

5.2 Synonym:

Presence of polysemous words in an expression increases semantic divergent nature. The replacement of a word with its synonym resonate the original sentence meaning without any alteration. A total of 531 variants (34.82% of the corpus) observe for 200 tweets while re-written by preserving meaning. For example:

Original tweet: A walk to remember is so amazing and inspiring.
Variant: A walk to remember is extremely astounding and motivating.

5.3 Idioms and Phrases:

Idiomatic expressions are a type of informal representation that has the same theme as the original expression with restrained meaning. Well-written text utilizes frequent use of idioms and phrases to make expression concise. Only 6 (0.4%) instances were found in the developed corpus where tweets were re-written using idioms and phrases. For example:

Original tweet: I seriously need a screen protector for my ipad
Variant: Need for screen protector for my ipad is dead serious.

5.4 Inter-clause re-orderings:

An expression may be divided into chunks known as information units to make expressions comprehensible. These information units are phonologically realized by the tone units (Rafajlovicova, 2002) (accessed November 7, 2016). Changing the order of tone units in an expression represents it in a different way. Communicatively, the most important positions in a clause/expression are the beginning and the end. Re-ordering of tone units helps to bring a tone unit in initial position or in end position to receive more focus. A total of 279 instances (18.29% of the total corpus) are there in the corpus under this feature. An expression with more number of tone units can be represented proportionally more ways by re-ordering clauses. For example:

Original tweet: a walk to remember is the only movie I like better than the book.
Variant 01: Better than the book I like the movie A Walk To Remember.
Variant 02: The only movie I like better than the book is A walk to remember.

5.5 Sentence Type Transformation:

In different writing style, a sentence may restate in different ways by changing the sentence structure. The work of (Mellon, 1969) (O'Hare, 1973) shows that the change in sentence structure is an effective method for improving content presentation without affecting the meaning of original text. A simple sentence can be transformed into a compound sentence by enlarging phrase or word into a coordinate clause or can be transformed into a complex sentence by enlarging a phrase into a subordinate clause such as Noun, Adjective or Adverb. A total of 198 number of instances (12.98%) are there in the corpus, formed by sentence structure transformations. For example:

Original tweet: check out the Yeti on Amazon it 's not too pricey.
Variant: Amazon is offering Yeti at a very cheap price. You can check it there.

5.6 Cleft:

Cleft sentences are used to focus on a particular part of the sentence to emphasize by introducing it with a kind of relative clause. Cleft construction breaks a sentence into two pieces of information in order to provide an extra focus on one piece. There are different types of clefts in English language (Calude, 2008) such as *if-because-cleft, it-cleft, wh-cleft, all-cleft, inferential-cleft and there-cleft*. The cleft structure involves important role in sentence content structure. Speakers/writers seek attention on salient parts of a message (Lambrecht, 2001) by using cleft based on highly linked content structure in the expression. Instances of tweet variants using cleft are less in number (1.25%) than other features in the develop corpus. For example:

Original tweet: I really hate when it rains on cinco de mayo.
Variant: ***When*** *it rains on cinco de mayo I really hate it.*

5.7 Addition or Deletion of non-impact words:

Natural Language like English becomes less ambiguous and more logical when it takes into account different unstated background assumption (Dewaele, 1999) while writing a sentence. Author / speaker may add or remove a word or more without altering its original meaning depending on the knowledge of context of that expression. This feature helps to resolve semantic ambiguity (Gorfein, 1989) (Grice, 1975) in the expression and makes it more clear, logical and self-possessed. A total of 235 numbers of variations (15.4%) exist in the corpus in this category. For example:

Original tweet: Rocky and john wall in that new quickaintfair Adidas commercial.
Variant: The new quickaintfair Adidas commercial was done by earlier model Rocky and John Wall.

Other than the above feature class, manual annotation process tag 173 number of sentences (11.34%) generated using two or more features (in table 1). For example, inter-clause reordering feature is clubbed with synonyms to restate the tweet here.

Original tweet: The only Nicholas Sparks movie I genuinely like is A Walk To Remember.
Variant: A Walk To Remember by Nicholas Sparks is one of my favorite film.

6 Measuring Thematic Closeness among Variants

We conduct an experiment to measure thematic closeness among the variants and tweets for each feature as in table 1. Our aim of this experiment is to measure the meaning intact status of tweet with its variants. In our experiment, we use three state-of-the-art semantic similarity measuring tools. First, the measure is (M_1) based on the work (Pilehvar et al., 2013). The tool measures similarity between the meanings of the words based on the sense of lexical items in the text by removing the ambiguity of word sense. Second measure (M_2) is based on the work (Pirró and Euzenat, 2010) which measures semantic similarity between ontology concepts of each expression. Equation 1 is used to calculate the similarity between tweet (concept c1) and its variant (concept c2), where msca(c1,c2) is the Most Specific Common Abstraction (msca) (Resnik, 1995) and Extended Information Content (eIC) characterize commonalities and differences.

$$sim_{FaITH}(c_1, c_2) = \frac{eIC(msca(c_1, c_2))}{eIC(c_1) + eIC(c_2) - eIC(msca(c_1, c_2))} \quad (1)$$

Third, the measure (M_3) is based on the method proposed by (Rus et al., 2013a) which uses Latent Semantic Analysis (LSA) model trained on the Wikipedia corpus from early January 2013 and the TASA corpus. The tool outperformed for measuring semantic similarity for paraphrase detection (Rus et al., 2013b).

Details of Average Similarity score of each tool for each feature class is reported in table 2.

Features	M_1	M_2	M_3
Activization/ passivization	0.924	0.946	0.962
Synonyms	0.960	0.907	0.940
Idioms & phrase	0.974	0.968	0.868
Inter-Clause re-ordering	0.913	0.965	0.885
Transforming sentence type	0.967	0.891	0.881
Cleft	0.906	0.899	0.926
Addition/deletion of non-impact words	0.907	0.893	0.969
Combined features	0.947	0.899	0.986

Table 2: Average semantic similarity score of each feature variants

7 Analysis and Discussion

In-depth analysis of semantic divergence of English tweets lead us to the following observations.

First, We observe that shorter length tweets have less variations in comparison to the longer length tweets. This is because longer tweets include more polysemous word as well as more clauses. For example the tweet *"Do not Amber Alert me"* has less semantic variants than the tweet *"I have to watch A Walk to Remember every time it shows"*. Second, thematic closeness of each feature class variants could not reach the maximum level due to two possible reasons. One, wrong spelling of words in tweets, which are corrected during re-writing. For example, the words *I'm* or *I m*, *r*, *NZ* in tweets are written as *I am*, *are*, *New Zealand* respectively in restated sentences. Two,

due to the limit of 140 characters, some tweets have incomplete word/clause at the end of tweet. The Rewritten form of that tweets exclude unwanted part. For example the tweet "That AAP Rocky Adidas commerical is hard af" includes incomplete word "af" .

8 Conclusion and Future Work

In this work we generated a corpus of 1525 semantically same sentences for 200 tweets by human annotators. This is the first corpus to represent semantic divergence of tweets. Our work explore the features for reinterpretation of a tweet meaning in different ways. Through human evaluation and experiments we justify meaning preservation in the variants.

We concentrated our current study on tweet corpus where a tweet can have a maximum of 140 characters only. This is an ongoing task. Our next target is to study the characteristics of semantic divergence of social media texts having un-restricted length as well as code-mixed social media texts.

References

A. Bell. 1995. Language style as audience design. In *Language in Society*. volume 13-2, pages 145–204.

Andreea S Calude. 2008. Demonstrative clefts and double cleft constructions in spontaneous spoken english. *Studia Linguistica* 62(1):78–118.

Schiling-Estes N Trudgill P Chambers J.K. 2006. *The handbook of language variation and change*. Blackwell.

Francis Heylighen & Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. In *Internal Report, Center "Leo Apostel", Free University of Brussels,*. MIT Press.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '08, pages 848–856.

David S. Gorfein. 1989. *Resolving semantic ambiguity / David S. Gorfein, editor*. Springer-Verlag New York.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts: Syntax and Semantics Volume 3*, Academic Press, New York, pages 41–58.

He He, Alvin Grissom II, John Morgan, Jordan L Boyd-Graber, and Hal Daumé III. 2015. Syntax-based rewriting for simultaneous machine translation. In *EMNLP*. pages 55–64.

Hideki Hirakawa, Kouichi Nomura, and Mariko Nakamura. 1994. An interactive rewriting tool for machine acceptable sentences. In *ANLP*. pages 207–208. http://aclweb.org/anthology-new/A/A94/A94-1043.pdf.

Jussi Karlgren. 1996. Stylistic variation in an information retrieval experiment. *CoRR* cmp-lg/9608003. http://arxiv.org/abs/cmp-lg/9608003.

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. *Computational Linguistics and Intelligent Text Processing* pages 446–457.

Knud Lambrecht. 2001. A framework for the analysis of cleft constructions. *Linguistics* 39(3; ISSU 373):463–516.

R.A. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 22(140):5–55.

Nitin Madnani and Bonnie J Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics* 36(3):341–387.

John C. Mellon. 1969. *Transformational sentence-combining: a method for enhancing the development of syntactic fluency in English composition [by] John C. Mellon*. National Council of Teachers of English Champaign, Ill.

Teruko Mitamura and Eric Nyberg. 2001. Automatic rewriting for controlled language translation. In *In Proceedings of the NLPRS 2002 Workshop on Automatic Paraphrasing: Theories and Applications*.

Frank. O'Hare. 1973. *Sentence combining; improving student writing without formal grammar instruction*. National Council of Teachers of English Urbana, Ill.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *ACL 2013*.

Giuseppe Pirró and Jérôme Euzenat. 2010. A feature and information theoretic framework for semantic similarity and relatedness. In *9th International Semantic Web Conference (ISWC2010)*. Springer, pages 615–630.

R. Rafajlovicova. 2002. Variation of clause patterns - reordering the information in a message. In `http://www.pulib.sk/elpub2/FHPV/Kacmarova1/pdf_doc/05.pdf`.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*. pages 448–453.

Stefan Riezler and Yi Liu. 2010. Query rewriting using monolingual statistical machine translation. *Computational Linguistics* 36(3):569–582.

Vasile Rus, Mihai C Lintean, Rajendra Banjade, Nobal B Niraula, and Dan Stefanescu. 2013a. Semilar: The semantic similarity toolkit. In *ACL (Conference System Demonstrations)*. pages 163–168.

Vasile Rus, Nobal Niraula, and Rajendra Banjade. 2013b. Similarity measures based on latent dirichlet allocation. In *Computational Linguistics and Intelligent Text Processing*, Springer, pages 459–470.

Graham Neubig Sakriani Sakti Tomoki Toda Tomoki Fujita and Satoshi Nakamura. 2013. Simple, lexicalized choice of translation timing for simultaneous speech translation. In *In Proceedings of Interspeech*.

Pidong Wang. 2013. *A Text Rewriting Decoder with Application to Machine Translation*. Ph.D. thesis, School of Computing, National University of Singapore, Singapore.