Duluth at SemEval-2017 Task 6: Language Models in Humor Detection

Xinru Yan
Department of Computer Science
University of Minnesota Duluth
Duluth, MN, 55812 USA
yanxx418@d.umn.edu

Ted Pedersen
Department of Computer Science
University of Minnesota Duluth
Duluth, MN, 55812 USA
tpederse@d.umn.edu

Abstract

This paper describes the Duluth system that participated in SemEval-2017 Task 6 #HashtagWars: Learning a Sense of Humor. The system completed Task A and Task B using ngram language models. This paper includes the results of our system with several post-evaluation runs. Our system is ranked highly in both tasks during evaluation.

1 Introduction

The following instructions are directed to authors of papers submitted to SemEval-2017 or accepted for publication in its proceedings. All authors are required to adhere to these specifications. Authors are required to provide a Portable Document Format (PDF) version of their papers. **The proceedings are designed for printing on A4 paper.**

SemEval papers should have the exact same format as ACL 2017 papers with two exceptions:

1. The review process is single-blind. **Submissions are not anonymous and should use the ACL camera-ready formatting**.

2. **Paper titles should follow the required template.** System description papers submitted by task participants should have a title of the form "[SystemName] at SemEval-2017 Task [TaskNumber]: [Insert Paper Title Here]". Task description papers submitted by task organizers should have a title of the form "SemEval-2017 Task [TaskNumber]: [Task Name]".

2 Background

Language models are a straightforward way to collect set of rules by utilizing the fact that words do not appear in an arbitrary order, which means we can learn a lot from a word and its neighbors (**?**). A statistical language model is a model that computes the probability of a sequence of words or an upcoming word (**?**).

Below are two examples of language modeling: To compute the probability of a sequence of words W given the sequence $(w_1, w_2, ...w_n)$, we have:

$$P(W) = P(w_1, w_2, ...w_n) \quad (1)$$

To compute the probability of an upcoming word $w3$ given the sequence (w_1, w_2), the language model gives us the following probability:

$$P(w_3|w_1, w_2) \quad (2)$$

The idea of word prediction with probabilistic models is called N-gram models, which predict the upcoming word from the previous N-1 words. An N-gram is a contiguous sequence of N words: a unigram is a single word, a bigram is a two-word sequence of words and a trigram is a three-word sequence of words. For example, in tweet "tears in Ramen #SingleLifeIn3Words", "tears", "in", "Ramen" and "#SingleLifeIn3Words" are unigrams; "tears in", "in Ramen" and "Ramen #SingleLifeIn3Words" are bigrams and "tears in Ramen" and "in Ramen #SingleLifeIn3Words" are trigrams.

In the study on how phrasing affects memorability, in order to analyze the characteristics of memorable quotes, researchers take language model approach to investigate distinctiveness feature and employ syntactic measures on the data to evaluate generality feature (**?**). Specifically, in favor of evaluating how distinctive a quote is, they evaluate its likelihood with the respect of the common language model which consists of the newswire sections of the Brown corpus. They employ six add-1 smoothed language modelsunigram, bigram, trigram word language models and unigram, bigram,

trigram Part of Speech (POS) language modelson the common languagemodel. The idea of using language models to assess the memorability of a quote is suitable for our purpose of detecting how humorous a twitter is. Except for using funny tweets provided by the task to train ngram language models, our system also trained ngram language models on English news data in order to evaluate how distictive, in this case, how funny, a tweet is comparing to news.
For our purpose, we trained unigrams, bigrams and trigrams on both sets of training data.

3 Method

Our system estimates tweet probability using ngram models. Specifically, it solves the given problem in four steps:

1. Corpus preparing and pre-processing: Collect all training data files to form one training corpus. Pre-processing includes filtering and tokenization.

2. Language model training: Build n-gram language models by feeding the corpus to KenLM Language Model Toolkit (**?**).

3. Tweet scoring: Get log probability for each tweet based on the trained ngram language model.

4. Tweet prediction: According to the log probability

    - Given two tweets, comparing two tweets and predicting which one is funnier (for subtask A)
    - Given a set of tweets associated with one specific hashtag, ranking tweets from the funnest to the least funny (for subtask B)

3.1 Corpus preparing and pre-processing

In our system, we used two distinct sets of training data: the tweets data and the news data. The tweets data is provided by the SemEval task. It includes 106 hashtag files, about 21580 tokens. In addition, we collected in total of 6.2 GB of English news data, about 2002655 tokens from the News Commentary Corpus and the News Craw Corpus from yeas of 2008, 2010 and 2011 [1].

[1] http://www.statmt.org/wmt11/translation-task.html#download

3.1.1 Preparing

To prepare the tweets data, the system takes in total of 106 hashtag files, which includes both trial_dir and train_dir from the task, and put all tweets in one plain text file to form the tweet training corpus. Each tweet is on its own line. Be aware that during the development phase of the system, we trained the language model on the train_dir data and tested it on the trial_dir data.
For the news data, the system reads in all the sentences from the news files and again, put them in one plain text file to form the news training corpus. Each sentence takes its own line.

3.1.2 Pre-processing

In general, the pre-processing consists of two steps: filtering and tokenization. The filtering step is mainly for the tweet training corpus. Also, we applied various filtering and tokenziation combinations on experiments to determine the best settings (see section 4).

- Filtering: the filtering process includes removing following elements from the :
    - URLs
    - Twitter user names with symbol @ indicating the user name
    - Hashtags with symbol # indicating the topic of the tweet

- Tokenization: For both training data sets we splitted text by space and punctuation marks

3.2 Language Model Training

Once we have the corpus ready, we use the KenLM Toolkit to train the n-gram language models on the corpus. Language models are estimated from the corpus using modified Kneser-Ney smoothing without pruning. KenLM reads in a plain text file and generates language models in arpa format. We trained three different language models – unigrams, bigrams and trigrams – for both training data sets. KenLM also implements back off technique, which simply means it applies the lower order ngram's probability along with its back-off weights if the ngram is not found. Instead of using the real probability of the ngram, KenLM applies base 10 logarithm scheme. Here is an example of the trigram model we trained on the tweets data:

ngram 1=21580
ngram 2=60624
ngram 3=73837
unigram:
-4.8225346 <unk>0
0 <s>-0.47505832
-1.4503417 </s>0
-4.415446 Donner -0.12937292
-3.4745252 Party -0.09994553
...

bigrams:
...
-0.9799023 Drilling Gulf -0.024524588
-3.9588327 of Mexicobe -0.024524588
...

trigrams:
...
-1.171928 I'll start thinking
-1.2377753 thinking he cares
...

Each ngram line starts with the base 10 logarithm probability of that ngram, followed by the ngram which consists of n words. The base 10 logarithm of the back-off weight for the ngram is followed after optionally. In this trigram language model trained on the tweets data, there are 73837 trigrams in total from the tweet training corpus. Notice that there are three "special" words in a language model: the beginning of a sentence denoted by <s>, the end of a sentence denoted by </s>and the out of vocabulary word denoted by <unk>. In order to be able to handle the unknown words to estimate the probability of a tweet more accurately, in all our experiments we keep the <unk>word in the language model. To figure out the best setting of language model for both tasks, we experiment using the language model with and without sentence boundaries.

3.3 Tweet Scoring

After training the ngram model, the next step is scoring. For each hashtag file that needs to be evaluated, based on the trained ngram language model, the system assigns a base 10 log probability for each tweet in the hashtag file. Here is an example of scored tweet from hashtag file Bad_Job_In_5_Words.tsv based on the ngram language model trained on the tweets data:
705511149970726912 The host of Singled Out #BadJobIn5Words @midnight -19.923433303833008 705538894415003648 Donut receipt maker and sorter #BadJobIn5Words @midnight -27.67446517944336

3.4 Tweet Prediction

The system sorts tweets for each hashtag file based on their score in descending order, meaning the most probable one is listed on the top. For Task A, given a hashtag file, the system goes through the sorted list of tweets, compare each pair of tweets and produces a tsv format file as the task asks for. For each tweet pair twee_1 and tweet_2, if tweet_1 has higher score, system outputs tweet_ids for the pair followed by "1" and followed by "0" otherwise. For Task B, given a hashtag file, the systm simply outputs tweet_ids in the order of the sorted list.

4 Experiments and Results

Includes result table

5 Discussion and Future Work

Includes using ML technique in the future

6 References

6.1 The Ruler

The ACL 2017 style defines a printed ruler which should be presented in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document without the provided style files, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment the `\aclfinalcopy` command in the document preamble.)

Reviewers: note that the ruler measurements do not align well with lines in the paper – this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. In most cases one would expect that the approximate location will be adequate, although you can also use fractional references (*e.g.*, the first paragraph on this page ends at mark 114.5).

6.2 Electronically-available resources

ACL provides this description in LaTeX2e (`acl2017.tex`) and PDF format (`acl2017.pdf`), along with the LaTeX2e style file used to format

it (`acl2017.sty`) and an ACL bibliography style (`acl_natbib.bst`) and example bibliography (`acl2017.bib`). These files are all available at `acl2017.org/index.php?article_id=9`. We strongly recommend the use of these style files, which have been appropriately tailored for the ACL 2017 proceedings.

6.3 Format of Electronic Manuscript

For the production of the electronic manuscript you must use Adobe's Portable Document Format (PDF). PDF files are usually produced from LaTeX using the *pdflatex* command. If your version of LaTeX produces Postscript files, you can convert these into PDF using *ps2pdf* or *dvipdf*. On Windows, you can also use Adobe Distiller to generate PDF.

Please make sure that your PDF file includes all the necessary fonts (especially tree diagrams, symbols, and fonts with Asian characters). When you print or create the PDF file, there is usually an option in your printer setup to include none, all or just non-standard fonts. Please make sure that you select the option of including ALL the fonts. **Before sending it, test your PDF by printing it from a computer different from the one where it was created.** Moreover, some word processors may generate very large PDF files, where each page is rendered as an image. Such images may reproduce poorly. In this case, try alternative ways to obtain the PDF. One way on some systems is to install a driver for a postscript printer, send your document to the printer specifying "Output to a file", then convert the file to PDF.

It is of utmost importance to specify the **A4 format** (21 cm x 29.7 cm) when formatting the paper. When working with `dvips`, for instance, one should specify `-t a4`. Or using the command `\special{papersize=210mm,297mm}` in the latex preamble (directly below the `\usepackage` commands). Then using `dvipdf` and/or `pdflatex` which would make it easier for some.

Print-outs of the PDF file on A4 paper should be identical to the hardcopy version. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs as soon as possible.

6.4 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact dimensions for a page on A4 paper are:

- Left and right margins: 2.5 cm
- Top margin: 2.5 cm
- Bottom margin: 2.5 cm
- Column width: 7.7 cm
- Column height: 24.7 cm
- Gap between columns: 0.6 cm

Papers should not be submitted on any other paper size. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs above as soon as possible.

6.5 Fonts

For reasons of uniformity, Adobe's **Times Roman** font should be used. In LaTeX2e this is accomplished by putting

```
\usepackage{times}
\usepackage{latexsym}
```

in the preamble. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX2e's default). Note that the latter is about 10% less dense than Adobe's Times Roman font.

Type of Text	Font Size	Style
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
the word "Abstract"	12 pt	bold
section titles	12 pt	bold
document text	11 pt	
captions	11 pt	
abstract text	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

6.6 The First Page

Center the title, author's name(s) and affiliation(s) across both columns. Do not use footnotes for affiliations. Do not include the paper ID number assigned during the submission process. Use the two-column format only when you begin the abstract.

Title: Place the title centered at the top of the first page, in a 15-point bold font. (For a complete guide to font sizes and styles, see Table 1) Long titles should be typed on two lines without a blank line intervening. Approximately, put the title at 2.5 cm from the top of the page, followed by a blank line, then the author's names(s), and the affiliation on the following line. Do not use only initials for given names (middle initials are allowed). Do not format surnames in all capitals (*e.g.*, use "Mitchell" not "MITCHELL"). Do not format title and section headings in all capitals as well except for proper names (such as "BLEU") that are conventionally in all capitals. The affiliation should contain the author's complete address, and if possible, an electronic mail address. Start the body of the first page 7.5 cm from the top of the page.

The title, author names and addresses should be completely identical to those entered to the electronical paper submission website in order to maintain the consistency of author information among all publications of the conference. If they are different, the publication chairs may resolve the difference without consulting with you; so it is in your own interest to double-check that the information is consistent.

Abstract: Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the paper by about 0.6 cm on each side. Center the word **Abstract** in a 12 point bold font above the body of the abstract. The abstract should be a concise summary of the general thesis and conclusions of the paper. It should be no longer than 200 words. The abstract text should be in 10 point font.

Text: Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document. Do not include page numbers.

Indent: When starting a new paragraph. Use 11 points for text and subsection headings, 12 points for section headings and 15 points for the title.

6.7 Sections

Headings: Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals. Do not number subsubsections.

Command	Output
{\"a}	ä
{\^e}	ê
{\`i}	ì
{\.I}	İ
{\o}	ø
{\'u}	ú
{\aa}	å

Command	Output
{\c c}	ç
{\u g}	ğ
{\l}	ł
{\~n}	ñ
{\H o}	ő
{\v r}	ř
{\ss}	ß

Table 2: Example commands for accented characters, to be used in, *e.g.*, BIBTEX names.

Citations: Citations within the text appear in parentheses as (**?**) or, if the author's name appears in the text itself, as Gusfield (**?**). Using the provided LATEX style, the former is accomplished using `\cite` and the latter with `\shortcite` or `\newcite`. Collapse multiple citations as in (**??**); this is accomplished with the provided style using commas within the `\cite` command, *e.g.*, `\cite{Gusfield:97,Aho:72}`. Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (**?**), but write as in (**?**) when more than two authors are involved. Collapse multiple citations as in (**??**). Also refrain from using full citations as sentence constituents.

We suggest that instead of

> "(**?**) showed that ..."

you use

> "Gusfield (**?**) showed that ..."

If you are using the provided LATEX and BibTEX style files, you can use the command `\citet` (cite in text) to get "author (year)" citations.

If the BibTEX file contains DOI fields, the paper title in the references section will appear as a hyperlink to the DOI, using the hyperref LATEX package. To disable the hyperref package, load the style file with the `nohyperref` option: `\usepackage[nohyperref]{acl2017}`

Digital Object Identifiers: As part of our work to make ACL materials more widely used and cited outside of our discipline, ACL has registered as a CrossRef member, as a registrant of Digital Object Identifiers (DOIs), the standard for registering permanent URNs for referencing scholarly materials. As of 2017, we are requiring all camera-ready references to contain the appropriate DOIs (or as a second resort, the hyperlinked ACL Anthology Identifier) to all cited works. Thus,

output	natbib	previous ACL style files
(**?**)	`\citep`	`\cite`
?	`\citet`	`\newcite`
(**?**)	`\citeyearpar`	`\shortcite`

Table 3: Citation commands supported by the style file. The citation style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

please ensure that you use BibTEXrecords that contain DOI or URLs for any of the ACL materials that you reference. Appropriate records should be found for most materials in the current ACL Anthology at `http://aclanthology.info/`.

As examples, we cite (**?**) to show you how papers with a DOI will appear in the bibliography. We cite (**?**) to show how papers without a DOI but with an ACL Anthology Identifier will appear in the bibliography.

As reviewing will be double-blind, the submitted version of the papers should not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, *e.g.*,

> "We previously showed (**?**) ..."

should be avoided. Instead, use citations such as

> "**?** (**?**) previously showed ... "

Please do not use anonymous citations and do not include acknowledgements when submitting your papers. Papers that do not conform to these requirements may be rejected without review.

References: Gather the full set of references together under the heading **References**; place the section before any Appendices, unless they contain references. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a citation as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (**?**). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the ACM *Computing Reviews* (**?**).

The LATEX and BibTEX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

Appendices: Appendices, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

6.8 Footnotes

Footnotes: Put footnotes at the bottom of the page and use 9 points text. They may be numbered or referred to by asterisks or other symbols.[2] Footnotes should be separated from the text by a line.[3]

6.9 Graphics

Illustrations: Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns. Color illustrations are discouraged, unless you have verified that they will be understandable when printed in black ink.

Captions: Provide a caption for every illustration; number each one sequentially in the form: "Figure 1. Caption of the Figure." "Table 1. Caption of the Table." Type the captions of the figures and tables below the body, using 11 point text.

6.10 Accessibility

In an effort to accommodate the color-blind (as well as those printing to paper), grayscale readability for all accepted papers will be encouraged. Color is not forbidden, but authors should ensure that tables and figures do not rely solely on color to convey critical distinctions. Here we give a simple criterion on your colored figures, if your paper has to be printed in black and white, then you must assure that every curves or points in your figures can be still clearly distinguished.

7 Translation of non-English Terms

It is also advised to supplement non-English characters and terms with appropriate transliterations and/or translations since not all readers understand all such characters and terms. Inline transliteration

[2] This is how a footnote should appear.

[3] Note the line separating the footnotes from the text.

or translation can be represented in the order of: original-form transliteration "translation".

8 Length of Submission

The ACL 2017 main conference accepts submissions of long papers and short papers. Long papers may consist of up to eight (8) pages of content plus unlimited pages for references. Upon acceptance, final versions of long papers will be given one additional page – up to nine (9) pages of content plus unlimited pages for references – so that reviewers' comments can be taken into account. Short papers may consist of up to four (4) pages of content, plus unlimited pages for references. Upon acceptance, short papers will be given five (5) pages in the proceedings and unlimited pages for references.

For both long and short papers, all illustrations and tables that are part of the main text must be accommodated within these page limits, observing the formatting instructions given in the present document. Supplementary material in the form of appendices does not count towards the page limit.

However, note that supplementary material should be supplementary (rather than central) to the paper, and that reviewers may ignore supplementary material when reviewing the paper (see Appendix A). Papers that do not conform to the specified length and formatting requirements are subject to be rejected without review.

Workshop chairs may have different rules for allowed length and whether supplemental material is welcome. As always, the respective call for papers is the authoritative source.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery* 28(1):114–133. https://doi.org/10.1145/322234.322243.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, and Jonand Lee Lillian Kleinberg. 2012. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 892–901. http://aclweb.org/anthology/P12-1094.

Daniël De Kok and Harm Brouwer. 2011. Natural language processing for the working programmer. Del.

Association for Computing Machinery. 1983. *Computing Reviews* 24(11):503–512.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. https://doi.org/10.18653/v1/P16-1001.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. http://aclweb.org/anthology/C14-1001.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pages 690–696. http://kheafield.com/professional/edinburgh/estimate_paper.pdf.

James H Martin and Daniel Jurafsky. 2000. Speech and language processing. volume 710.

A Supplemental Material

ACL 2017 also encourages the submission of supplementary material to report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.

Nonetheless, supplementary material should be supplementary (rather than central) to the paper. **Submissions that misuse the supplementary material may be rejected without review.**

Essentially, supplementary material may include explanations or details of proofs or derivations that do not fit into the paper, lists of features or feature templates, sample inputs and outputs for a system, pseudo-code or source code, and data. (Source code and data should be separate uploads, rather than part of the paper).

The paper should not rely on the supplementary material: while the paper may refer to and cite the supplementary material and the supplementary material will be available to the reviewers, they will not be asked to review the supplementary material.

Appendices (*i.e.* supplementary material in the form of proofs, tables, or pseudo-code) should come after the references, as shown here. Use `\appendix` before any appendix section to switch the section numbering over to letters.

B Multiple Appendices

. . . can be gotten by using more than one section. We hope you won't need that.