ID: 64

Proceedings of the International Workshop
on Modern Science and Technology
Kitami, Japan September 2010

Improvements in an Experimental Annotated Linguistic Pre-processing System of Casual English

Eleanor Clark and Kenji Araki

Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan
{eleanor, araki}@media.eng.hokudai.ac.jp

Abstract: We present functional improvements and the results of an evaluation experiment in a text processing system, CECS (Casual English Conversion System). The purpose of CECS is to normalize the casual, error-ridden English that is frequently a feature of new communication media including Internet message boards, blogs, emails, chat applications, cellphone SMS messages, and services such as Twitter, into regular English.

This paper firstly describes recent improvements made to the system by the implementation of a phrase matching capability using a trie-type algorithm. This allows far more database coverage of slang phrases and common abbreviations than the previous revision, which used a word matching algorithm, and also facilitates progress towards handling word-sense disambiguation (WSD) problems.

The results of two evaluation experiments are also discussed in detail. The system has been tested for usability as a pre-processing tool for Machine Translation (MT), as well as preliminary user-based human evaluation experiments. Both experiments show promising results, with a sharp decrease in non-translated words in the MT experiment, and a significant increase in self-assessed reader comprehension in the human evaluation experiment.

Keywords: Natural Language Processing, Machine Translation, Text Normalization, Text Cleaning, Blogs, Twitter

1. Introduction

The rapid expansion of Internet use, electronic communication and user-oriented media such as social networking sites, blogs and microblogging services has led to an exponential increase in the need to understand casual written English, which often does not conform to rules of spelling, grammar and punctuation. Despite this, text normalization is commonly seen as a "messy chore" [1], and remains a somewhat niche topic of research. Studies which attempt to tackle this problem generally use a fully automated, statistical approach [2,3][1]; however, we propose that a combination of automated and manual techniques is a potentially more useful approach to this problem. Accordingly, our aim is to develop a method which uses automated tokenization, word matching and replacement techniques in combination with a high-quality, large scale, manually compiled database. We present recent progress on this system, CECS (Casual English Conversion System).

CECS has two applications: as pre-processing on noisy input for automated Natural Language Processing tasks such as Machine Translation or Information Retrieval; and as a standalone system for human users, to aid non-native speakers' reading comprehension of informal written English, the irregularity of which may pose a barrier to their positive participation in 21st Century international communications.

This user-oriented educational aspect of CECS is complemented by the inclusion of annotation on linguistic and/or cultural aspects of each word or phrase converted by the system. At present, the system's knowledge base for text replacement is a manually compiled database of 912 items, although expansion of the database is constant and regular.

2. CECS: Casual English Conversion System

2.1 System Overview

The process of CECS is shown schematically in Figure 1.

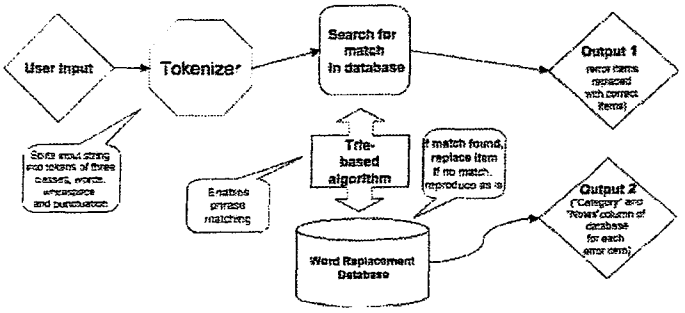

Figure 1: Process of CECS

CECS is written in the Python programming language. Firstly, user input is tokenized using a strictly regular grammar defined in PyParsing[2], which defines words and punctuation as separate tokens, and allows combinations. "Main characters" are defined as the letters from a-z and A-Z, numbers 0-9 (in case of spellings which incorporate numbers such as "*gr8*" for "*great*"), and selected punctuation marks which may appear mid-word such as apostrophe ("*don't*"), hyphen

[1] As this main goal of this paper is to introduce the results of recent experiments, a detailed review of related works is beyond the limits of space. Discussion of a wide range of relevant research can be found in the previous paper [4].

[2] *http://pyparsing.wikispaces.com/*

("*mid-word*"), and asterisk for censor avoidance spellings ("*s****"), etc. "Other characters" are defined as all other ASCII characters, and whitespace and carriage returns are defined separately. A token is thus defined here as either a word composed of main characters ("English word") or composed of other characters ("punctuation token").

Tokenized input is then passed through the database to find a match, using a trie-type data structure. When a match is found, the normalized English equivalent is displayed in the user interface in the "Output" pane, and the replaced item's category and notes, where present, are displayed in the "Notes" pane. Tokens not found in the database are passed through unchanged.

2.2 Recent Improvements

In addition to substantial database expansion since the first version of the system [5], the main recent improvement made to the process of CECS has been the addition of a trie data structure (Figure 2) to enable phrase matching. The first version of CECS used a simple word-for-word replacement algorithm, which, while allowing several hundred vocabulary items to be normalized, was badly limited in not permitting phrases of two or more words to be matched in the database. This prompted a need to revise the process.

In the new version, the database is recursively loaded into a trie to allow easy item lookup, tokenized by the same tokenizer used for input. Database entries which are a front-anchored substring are allowed, but full matches are not. Using this data structure, multi-word phrase matching is enabled. Text processing is reasonably fast: a two thousand word text is fully searched, matched and replaced in less that one second.

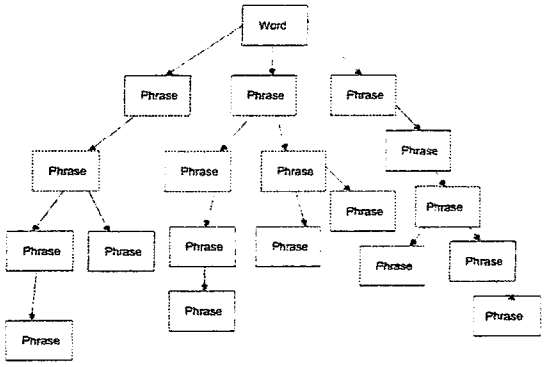

Figure 2: Representation of a trie data structure

The benefits of implementing phrase matching into CECS are significant. Firstly, slang phrases constituting more than one word can be matched in the database; secondly, problems regarding word sense disambiguation (WSD) problems within the sphere of casual English usage can be tackled for the first time in this system. When a word exists as a regular English word but is often used in casual English to mean something else, it previously could not be entered in the database as a single item, in case the regular meaning of the word was being used. As an example, the regular English word "bout" is commonly used as a shortened form of "about". However, it may also be used in its original meaning, as the example sentences below show.

> **Regular usage:** The semi-final **bout** between France and Holland was disappointing.
>
> **Casual usage**: i dont know **bout** u, but i deffo want 2 c da footie game 2nite. (*I don't know about you, but I definitely want to see the football game tonight*)

As well as being confusing to non-native readers, this word causes problems to MT applications, which tend to translate it in its original meaning of "game" or "battle", rendering many casual English sentences difficult to understand after translation. With phrase matching in CECS, common combinations of "bout" which can *only* be used in the sense of "about" can be added into the database. Thus, pre-processing casual English with CECS can improve MT handling of such vocabulary items. Below is a section of the database entries containing "bout":

Input	Normalization
bout dis	about this
bout her	about her
bout him	about him
bout it	about it
bout that	about that
bout this	about this
bout me	about me
bout you	about you
care bout	care about
hear bout	hear about
nothing bout	nothing about
nuffing bout	nothing about
kno bout	know about
know bout	know about

This approach also proves useful for normalizing numbers which have been used as phonetic substitutions, e.g. "4" for "for", "2" for "to" or "too", etc. It would have been extremely inaccurate in the previous version of the system to automatically convert all instances of the number "4" to "for", however, in this version it is possible to convert a high number of occurrences correctly using carefully designed combinations. Thus, we can define the rules for the usage of these items manually, and automatically convert appropriately with CECS.

While this strategy of addressing WSD may not yet cover every potential possibility and usage, it is logical that the combinations used are finite and thus can be entered in the database. As more data is collected, analyzed and more examples gathered, the quality and coverage of the database further increases. However, if database size passes a certain point, there is a risk of simply becoming a vast collection of separate examples, rather than rules. Accordingly, in future work on CECS, we plan to create a high-quality,

effective database within the limits of around three thousand entries.

3. Evaluation Experiment A: CECS Output for MT Use

3.1 Experiment Overview

One hundred sentences from the popular microblogging service Twitter[3] were run through two well-known free MT applications, Google Translate[4] and Systran[5]. The same sentences were then pre-processed with CECS and run through Google Translate and Systran a second time. The quality of the resulting translations was compared by measuring error incidence. The working language pair used was English to Japanese. Compared to a previous preliminary experiment on CECS [4], this was a more detailed experiment with a bigger data set and two clearly defined categories of error.

Of the 100 Twitter sentences, 20 were "known" sentences, i.e., they had been analyzed for error words and those items were pre-entered into the database. The remaining 80 were "unknown" sentences.

The method to assess error incidence was as followed. MT errors were counted manually in two separate categories, "non-translated word" ("NTW") and "wrongly translated word" ("WTW"). An NTW is defined here as the MT application simply reproducing an item as roman letters or numbers, and not converting to Japanese at all. Numbers are only considered to be NTWs if they are used specifically as phonetic replacements for words (e.g *"I luv u 2"* for *"I love you too"*.) A WTW is defined as a Japanese word that is completely semantic different from the English meaning. To illustrate, in the translated sentence below NTWs are underlined and WTWs are shown in bold text.

Raw input: now y would u say sucha thing?! I make SURE 2 keep a clean house...lol!

Google MT Result :
今yはuは言うスハのこと！？ 私は、第2保持する家を掃除する...笑！

3.2 Experiment Data

The sentences used in Experiment A were gathered from a 10.5 million "tweet" (Twitter posting) Twitter corpus as compiled and publicly released by Choudhury [6]. The corpus contains tweets from 200,000 unique users collected between 2006 and 2009; the 100 sentences used in our experiment were taken from the September 2009 section of the corpus. The user tweets used as data for this experiment were essentially selected at random, but with the following criteria: a) the sentence is written entirely in English b) the sentence contains at least two "errors" or non-dictionary words for CECS to be tested on. Accordingly, non-English and grammatically perfect sentences were discarded.

Sentences longer than 30 words were split into two data items. Any Twitter usernames and linked URLs were removed from the sentences prior to use in the experiment.

Average sentence length was 15.35 words for raw input, and 15.99 words after pre-processing with CECS. The slight increase is due to the fact that some phrases are expanded from contractions or acronyms, e.g. *omg* becoming *Oh my God*, *wassup* becoming *What's up*, etc.

3.3 Experiment Results

The results of Evaluation Experiment A for all sentences are summarized in Table 1. Table 2 shows results only for "known" sentences (training data). Table 3 shows results only for "unknown" sentences (test data).

Table 1: Error counts in all sentences (100)

	Raw Input		CECS Output	
	NTWs*	WTWs	NTWs	WTWs
Google MT	2.78	1.55	0.83	0.86
Systran MT	3.83	0.84	0.77	0.56
Avg. of both MT systems:	**3.31**	**1.2**	**0.8**	**0.71**

*All NTW (non-translated word) and WTW (wrongly translated word) counts are given as an average per sentence.

Table 2: Error counts in known sentences (20)

	Raw Input		CECS Output	
	NTWs*	WTWs	NTWs	WTWs
Google MT	2.65	1.5	0.5	0.55
Systran MT	3.65	0.8	0.65	0.7
Avg. of both MT systems:	**3.15**	**1.15**	**0.58**	**0.63**

*All NTW (non-translated word) and WTW (wrongly translated word) counts are given as an average per sentence.

Table 3: Error counts in unknown sentences (80)

	Raw Input		CECS Output	
	NTWs*	WTWs	NTWs	WTWs
Google MT	2.81	1.56	0.91	0.93
Systran MT	3.87	0.85	0.8	0.52
Avg. of both MT systems:	**3.34**	**1.21**	**0.86**	**0.76**

*All NTW (non-translated word) and WTW (wrongly translated word) counts are given as an average per sentence.

[3] www.twitter.com
[4] http://translate.google.com
[5] http://www.systranet.com

As seen in all three tables, while the decrease in non-translated words after pre-processing with CECS is sharp, the decrease in wrongly translated words is much less significant. It can be suggested that a sizeable proportion of wrongly translated words are of regular dictionary words that have been mistranslated due to Google Translate and Systran's currently limited handling of WSD problems, as discussed in Section 3.4, and are not target input for CECS. However, the decrease in non-translated words is highly notable, and this experiment has proven CECS to be useful in reducing the amount of non-translated words in English to Japanese machine translation of casual language.

Comparison between Table 2 and Table 3 reveals that prior entry in the database dramatically increases accuracy, as would be expected. However, the decrease in NTWs in Table 3 from 3.34 to 0.86, a significant drop, shows that CECS' current level of database coverage gives reasonable performance. As the database is constantly updated, this is expected to increase.

3.4 Discussion: Error Analysis

As can be seen in Tables 1 to 3, the two MT systems had different error handling. On the whole, Systran's incidence of NTWs in raw input was significantly higher than Google Translate's, but dropped to be slightly lower than Google Translate's after pre-processing with CECS. Incidence of WTWs was lower in Systran for both raw input and system output. It could be observed from this experiment's results that CECS is more effective as a pre-processing tool for Systran, a rule-based MT application, than for Google Translate, a statistical MT application; at least, for the language pair English to Japanese.

In terms of causes for errors in MT output, a significant number of NTWs resulted from casual or error vocabulary items which were absent from the database. In fact, a small number of sentences were entirely unchanged after being passed through CECS, as all error items contained were out-of-database. Although some of these items were slang or abbreviations (later added to the database after the experiment was concluded), many were typing or spelling errors. This last problem could be better addressed by integrating an open source spellchecker, such as GNU Aspell[6], into a future version of CECS.

Another cause of errors, again mainly in the NTW category, was the occurrence of named entities in the input such as a person's name, product, organization or other body. Famous place names and common Western names were usually translated into correct *Katakana*[7] equivalents, but nicknames, non-English or rare English names, as well as most products and brand names often appeared as NTWs. A possible solution to this would be the future implementation of a Web mining function, searching large knowledge collections such as Wikipedia[8] for definition of the named entity. This could be given as a linked URL in the system output.

The main cause of WTWs was WSD problems, where a word with more than one meaning was translated incorrectly for the context. This occurred both with casual English vocabulary items and regular dictionary English words, indicating that Google Translate and Systran are currently not fully able to tackle the complex issue of WSD.

An example of a successfully normalized sentence from Experiment A is shown below. NTWs are immediately obvious even to a non-Japanese reader.

Raw input: 4 yr old went 2 her first funeral. Asked me "Grandma when U die, will U invite me 2 UR funeral & can I sit on the front seat?" I said "yes"

Google Translate: 4 yrは、古い2彼女の最初の葬儀を行った。私質問"おばあちゃんは、Uが死ぬ、Uは2ウルの葬儀&Iは、前の座席に座ってすることができます私を招待のだろうか？"私は言った"はい"
NTWs: 7 (yr, 2, U, U, 2, &, I)
WTWs: 1 (ウル from input "UR")

Systran : 4 yr oldは2彼女の最初葬式行きました。 Uが死ぬ場合私に「祖母頼まれて、Uは私を2 URの葬式誘いましたり及び前の座席で置かれることができます私か」。を 私は「はい」言いました
NTWs: 7 (yr, old, 2, U, U, 2, UR) WTWs: 0

System output: 4 year old went to her first funeral. Asked me "Grandma when you die, will you invite me to your funeral and can I sit on the front seat?" I said "yes"

Google Translate: 4歳の彼女の最初の葬儀に行きました。私"おばあちゃんが死んで、あなたの葬式に、私を招待し、私は前の座席に座ることができるか？"私は言った"はい"
NTWs: 0 WTWs: 0

Systran: 4歳児は彼女の最初葬式に行きました。 死ぬ場合私に「祖母頼まれて、あなたの葬式に私を誘い、前の座席で置かれることができます私か」。を 私は「はい」言いました
NTWs: 0 WTWs: 0

As this sentence used only casual vocabulary items already in the database, NTW occurrence was reduced significantly. WTW incidence was originally low or non-existent. Note that the numbers were normalized correctly: "4" was unchanged, as it referred to the age of four, but "2" was changed to "to" using the database entry "2 ur" = "to your".

4. Evaluation Experiment B: CECS Output for English Learners

4.1 Experiment Overview

Six days after Experiment A was completed, a second experiment was conducted over a period of five days. Experiment B is the first attempt to assess CECS with human evaluators. Ten learners of English between the

[6] http://aspell.net/
[7] Japanese syllabic alphabet, primarily used for foreign words.
[8] www.wikipedia.org

ages of 23 and 64 (9 evaluators were Japanese, 1 was Chinese) completed two questionnaires, in which they were asked to assess their understanding of 20 sentences. The first questionnaire used raw input for the sentences, and the second questionnaire used the same sentences after processing by CECS. No participants were allowed to see the corrected sentences until they had submitted the first questionnaire.

Rankings were made on a five-point semantic differential scale, as follows:

> **Question:** How much of the sentence can you understand?
> **1.** None at all **2.** A little **3.** Some **4.** Most **5.** All

Evaluators were also asked to give a reason for why they could not understand part or all of each sentence. They were given three choices: vocabulary, grammar and context. Attributing more than one reason to failing to understand a sentence was possible.

Evaluators were also asked to assess their level of English comprehension on a scale of 1 (very basic) to 5 (excellent).

4. 2 Experiment Data

The 20 sentences used in the human evaluation experiment were taken from Experiment A, unchanged. They were selected randomly from the group of 80 "unknown" sentences, in order to test database coverage objectively.

4.3 Experiment Results

First, the participants' English comprehension self evaluation results were as follows. Three people rated themselves as 2 (Basic), four people rated themselves as 3 (Fair), and three people rated themselves as 4 (Good). None of the evaluators rated their English at either extreme end of the scale (1 and 5).

Overall, average understanding of the 20 sentences increased by exactly one semantic differential point: evaluator comprehension of the sentences averaged at 2.89 for raw input, on the low side of "Some" on the semantic scale, and 3.89 for system output, or slightly lower than "Most" on the semantic scale. As the evaluator's level of English reading ability could be seen to have a strong influence on comprehension levels, Tables 4 shows the results grouped by participants' self-assessed English comprehension level.

Table 4

	Level 2 English (Basic)	Level 3 English (Fair)	Level 4 English (Good)
Reader understanding: **Raw input**	1.53*	3.26	3.88
Reader understanding: **System output**	2.93	4.21	4.53

*Reader understanding is given as an average of answers made on a semantic differential scale of 1-5, where 5 is full comprehension.

Table 4 shows a clear increase in sentence comprehension at all levels, but the difference is most significant at the Basic level. This fact suggests that CECS may be most useful for lower-intermediate learners of English.

When understanding was rated as less than 5, evaluators were asked to give reason(s). Figures 3 and 4 show the proportion of all reasons given by all participants. Note that participants gave two or three reasons for some sentences, and one or none for others.

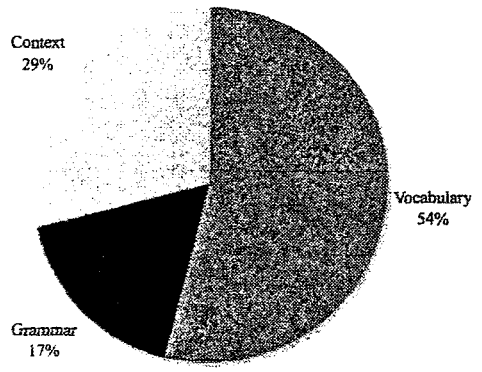

Figure 3: Reasons for Non-comprehension (Raw Input)

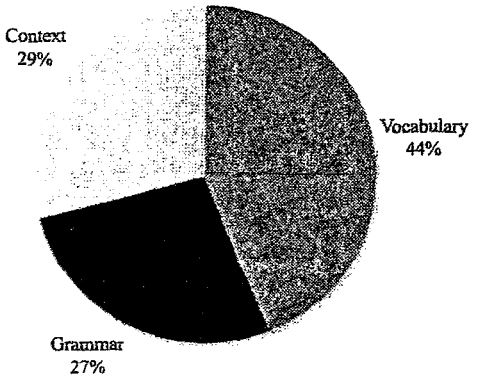

Figure 4: Reasons for Non-comprehension (System Output)

While vocabulary was the main problem in over half of cases in raw input sentences, this proportion was slightly reduced in system output sentences. The proportions do not show a drastic change, although the balance between all three causes for non-comprehension is more equal in the system output results. Vocabulary remains the main reason for non-comprehension; however, this may be attributed to non-comprehension of difficult "regular" English words as well as remaining error words in sentences not corrected by the system. It should be noted that Figure 3 shows the proportions of 228 separate participant answers, whereas Figure 4 is of 146 answers. This drop reflects the fact that the number of reasons given for non-comprehension fell as comprehension increased.

4.4 Error and Human Feedback Analysis

Several sentences were not completely normalized, as the sample came from unknown data: many error items were not in CECS database. An example of a partially corrected sentence from Experiment B is as follows:

> **Raw input:** Gr8 ldrs surround themselves w/others who compensate 4 their weeknesses. Who r u surrounded by?
>
> **System output:** Great ldrs surround themselves with others who compensate 4 their weeknesses. Who are you surrounded by?

Due to the fact that some vocabulary items, particularly "ldrs" (leaders) which is the subject of the first sentence, were not converted, several evaluators assigned a low score to this sentence even after pre-processing with CECS. An example of a more successful conversion is as follows:

> **Raw input:** B4 u run, u need 2 walk, b4 walking u need 2 crawl
>
> **System output:** before you run, you need to walk, before walking you need to crawl

This sentence, which received low scores in raw input form – mostly attributed to vocabulary by participants, probably due to the heavy use of numerical substitutions – gained a high proportion of "4" and "5" scores after pre-processing with CECS.

In order to provide a channel for evaluator feedback, a free comments box was given at the end of the questionnaire. Notable comments are shown below (two comments, indicated in italics, have been translated from their original Japanese) :

> I think it is very difficult for the English learners as a second language to understand them. But I felt like being a detective to solve those problems.
>
> I do not usually use the computer English, so. at first it was harder for me to understand the sentences, but gradually I was able guess the meaning of sign language such as 2, 4 , ur, and others.
>
> *When there are missing words, it is very difficult to understand the sentence, especially if the subject is missing.*
>
> *Without proper capitalization and punctuation, I can't see where the clauses start and end. It's a mess.*
>
> There are a few words I cannot understand. ldrs , nyhw and xo. But I understood the sentences much better than in the previous questionnaire. (In relation to Part 2)

Comments from evaluators were illuminating in terms of identifying future areas of improvement for CECS. Particularly, the problem of missing punctuation and words is a non-trivial issue, which also caused significant problems to the two MT systems seen in Experiment A. It can be considered that using the phrase matching capability to address common vocabulary omissions would be useful in forthcoming updates to the CECS database.

5. Conclusion

We have presented recent improvements made to a text normalization system, and the results of two experiments. Both the Machine Translation-based experiment and human evaluation-based experiment showed positive results, with a significant reduction in non-translated words in the former, and a notable improvement in reader comprehension in the latter after pre-processing Twitter sentences with our system. Human evaluator feedback emphasized both the usefulness and need for this system, and gave us ideas for future improvements.

We consider that the main tasks hereafter will be the ongoing expansion of the database, and developing the system with additional techniques such as the integration of an open-source spellchecking tool for dealing with a wider range of spelling errors, and the implementation of a Web mining algorithm for access to a wider knowledge base.

References

[1] Sproat, R., A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 2001. Normalization of non-standard words. In *Computer Speech and Language, 15(3)* pp287-333

[2] Aw, A., M. Zhang, Z.Z Fan, P.K. Yeo and J. Su. 2006. A Phrase-based Statistical Model for SMS Text Normalization. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Session* pp33-40

[3] Henriquez, C.A and A. Hernandez, A Ngram-based Statistical Machine Translation Approach for Text Normalization on Chat-speak Style Communications, *Proceedings of CAW2.0 2009* Madrid, 2009 pp. 1-5

[4] Clark, E, T. Roberts and K. Araki, Towards a Pre-processing System for Casual English Annotated with Linguistic and Cultural Information, *Proceedings of Computational Intelligence 2010*, Hawaii, 2010, in print

[5] Clark, E and K. Araki, A Basic Annotated Linguistic Pre-processing System for Casual English with Ideas for Expansion, *Proceedings of GCOE-NGIT 2010*, Sapporo, 2010, pp. 184-185

[6] Choudhury, M. D., Lin, Y.R., Sundaram, H., Candan, K. S., Xie, L., Kelliher, A., How Does the Sampling Strategy Impact the Discovery of Information Diffusion in Social Media? *Proceedings of the 4th Int'l AAAI Conference on Weblogs and Social Media*, Washington DC, 2010, in print