Volume 4 ▪ Issue 1 ▪ August 2013

INTERNATIONAL JOURNAL OF

COMPUTATIONAL LINGUISTICS (IJCL)

Publication Frequency: 6 Issues / Year

ISSN : 2180-1266

CSC PUBLISHERS
http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTATIONAL LINGUISTICS (IJCL)

VOLUME 4, ISSUE 1, 2013

EDITED BY
DR. NABEEL TAHIR

ISSN (Online): 2180 - 1266

International Journal of Computational Linguistics (IJCL) is published both in traditional paper form and in Internet. This journal is published at the website http://www.cscjournals.org, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJCL Journal is a part of CSC Publishers
Computer Science Journals
http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTATIONAL LINGUISTICS (IJCL)

Book: Volume 4, Issue 1, August 2013
Publishing Date: 30-08-2013
ISSN (Online): 2180-1266

IJCL Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers, 2013

EDITORIAL PREFACE

The International Journal of Computational Linguistics (IJCL) is an effective medium for interchange of high quality theoretical and applied research in Computational Linguistics from theoretical research to application development. This is the *First* Issue of Volume *Four* of IJCL. The Journal is published bi-monthly, with papers being peer reviewed to high international standards. International Journal of Computational Linguistics (IJCL) publish papers that describe state of the art techniques, scientific research studies and results in computational linguistics in general but on theoretical linguistics, psycholinguistics, natural language processing, grammatical inference, machine learning and cognitive science computational models of linguistic theorizing: standard and enriched context free models, principles and parameters models, optimality theory and researchers working within the minimalist program, and other approaches.

IJCL give an opportunity to scientists, researchers, and vendors from different disciplines of Artificial Intelligence to share the ideas, identify problems, investigate relevant issues, share common interests, explore new approaches, and initiate possible collaborative research and system development. This journal is helpful for the researchers and R&D engineers, scientists all those persons who are involve in Computational Linguistics.

Highly professional scholars give their efforts, valuable time, expertise and motivation to IJCL as Editorial board members. All submissions are evaluated by the International Editorial Board. The International Editorial Board ensures that significant developments in image processing from around the world are reflected in the IJCL publications.

IJCL editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Scribd, CiteSeerX Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCL. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJCL provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Computational Linguistics (IJCL)

EDITORIAL BOARD

TABLE OF CONTENTS

Volume 4, Issue 1, August 2013

Pages

Arabic SentiWordNet in Relation to SentiWordNet 3.0

Samah Alhazmi *alhazmis@cs.man.ac.uk*
Text Mining Research Group
School of Computer Science
University of Manchester
Manchester, UK, M13 9PL

William Black *william.black@manchester.ac.uk*
School of Computer Science
& National Centre for Text Mining
Manchester Institute of Biotechnology
University of Manchester
Manchester, UK, M1 7DN

John McNaught *John.McNaught@manchester.ac.uk*
School of Computer Science
& National Centre for Text Mining
Manchester Institute of Biotechnology
University of Manchester
Manchester, UK, M1 7DN

Abstract

Sentiment analysis and opinion mining are the tasks of identifying positive or negative opinions and emotions from pieces of text. The SentiWordNet (SWN*)* plays an important role in extracting opinions from texts. It is a publicly available sentiment measuring tool used in sentiment classification and opinion mining. We firstly discuss the development of the English SWN for versions 1.0 and 3.0. This is to provide the basis for developing an equivalent SWN for the Arabic language through a mapping to the latest version of the English SWN 3.0. We also discuss the construction of an annotated sentiment corpus for Arabic and its relationship to the Arabic SWN.

Keywords: Opinion Mining, Sentiment Analysis, WordNet, SentiWordNet, Arabic.

1. INTRODUCTION

Text mining involves the automated extraction of information from texts, often from large volumes of texts. An area of growing interest within text mining is opinion mining, which involves assessing whether a text is objective or subjective, and, if it is subjective, whether it is positive or negative [1, 2, 3, 4]. This is relevant to many tasks such as determining public opinion about a particular product, or tracking movements in public opinion in relation to questions of public policy. It involves both determining the polarity of a text (if any) and the strength of the text polarity. Texts may be weakly, mildly or strongly positive or negative and these differences can be highly relevant to the conclusions that can be drawn from the analysis.

Given the intrinsically subjective nature of opinions, assessing the quality of the results generated by any tool raises particular difficulties [1, 5].

This paper sheds light on the development of SentiWordNet (SWN) 1.0 and 3.0, a publicly available resource used in sentiment classification and opinion mining [1], and describes and compares the effectiveness of the English versions. SWN is an evolving resource that maps to consecutive versions of WordNet (WN) [15] — a resource consisting of synsets[1] that list (disambiguate) the multiple senses of words which are often ambiguous. This means that they may have multiple senses and vary in meaning by context. These disambiguated words are then glossed (explained), and it should be noted that the words in the gloss are not themselves disambiguated in the original WN.

SWN can also be applied across different languages but, as it maps to WN, it requires an appropriate version of WN for the language in question.

In the case of Arabic, this requires the development of an Arabic version of WN 3.0 that is then mapped to the English WN 3.0. In the absence of this, the development of sentiment analysis for Arabic texts will fall behind in one of the most promising areas for text mining. What is more, the existing Arabic WN 2.0 does not contain an extensive range of synsets and this inadequacy needs to be addressed rapidly. Based on this fact, this paper contributes toward presenting an Arabic SWN in relation to the latest version of the English SWN 3.0, taking into account upgrading the Arabic WN 2.0 to version 3.0.

Section 2 gives a brief summary of lexicons that are available for sentiment. Section 3 draws a brief history of SWN and discusses the English versions 1.0 and 3.0. Section 4 then outlines the preparatory steps that need to be taken to develop a version of SWN to support analysis of Arabic texts, and includes a discussion of the mapping process between the English and the Arabic versions. Section 5 discusses related work. Finally, future directions are briefly considered in the conclusion, section 6.

2. BACKGROUND

Numerous lexicons are available for sentiment analysis. There can be low or quite significant degrees of disagreement among them. It is possible to intentionally use such contrasts in solving conflicts or, alternatively, they may be accepted to be genuine areas of uncertainty.

WordNet may be utilised to determine useful lexicons from small seed sets, including in cases where the differences are not clearly encoded within WordNet.

One of the key benefits of lexical induction is the ability to include domain-specific effects. Table 1 summarises a number of lexicons and their characteristics.

Lexicons	Brief Descriptions
SentiWordNet [1, 5, 15]	SentiWordNet assigns positive or negative numerical sentiment values to WordNet synsets. It is available for free as long as it is not used for financial gain, with business users required to seek a license.
Liu's Opinion Lexicon [22, 23]	It provides for spelling errors, morphological variation, vernacular and internet terminology.
MPQA Subjectivity Lexicon [24]	The Multi-Perspective Question Answering (MPQA) Subjectivity Lexicon is distributed under the GNU Public License.
Harvard General Inquirer [25]	The Harvard General Inquirer is a lexicon which assigns syntactic, semantic as well as practical data to part-of-speech tagged words.
LIWC [26]	Linguistic Inquiry and Word Counts (LIWC) is a proprietary database of classified common terms. It is priced at $90 and has categories which are very close to the ones in the Harvard General Inquirer.

TABLE 1: Summary of number of lexicons and their characteristics.

[1] Synsets are sets of "senses" for a word that clarify the particular sense in which it can be used. The words that clarify each sense are called the "gloss".

The aforementioned lexicons have only basic divergence in categorisation. They possess word banks that are not the same and, thus, can only be contrasted to a limited extent.

For the purposes of this paper, we have chosen SentiWordNet as a sentiment lexicon for our work, to build an Arabic version of this lexicon, as will be discussed later.

3. SENTIWORDNET: BRIEF HISTORY

Researchers have attempted to develop systems to automatically label words that indicate opinions as being either positive or negative [6,7,8]. The prior and related question of whether a word is in fact a marker of opinion or not, whether it is subjective or objective, has received less attention [9]. Early attempts involved labelling words without making distinctions between the different senses in which a word may be used, with the result that the word rather than its sense is classified. This has limitations as the same word very often has multiple senses and any system that fails to capture these variations in meaning is severely restricted in functionality and reliability.

2.1 SentiWordNet 1.0

Esuli and Sebastiani [1] attempted to address this limitation by developing a resource, SWN (version 1.0), that goes beyond simply listing and classifying a word to classify the sense in which it is being used. That is, one word may — and usually does — have multiple senses. In the WN world, such multiple senses for a word are called synsets.

In SWN 1.0, Esuli and Sebastiani augmented each WN synset by assigning a numerical score in three categories: Obj (Objective), Pos (Positive) and Neg (Negative). These mean that a synset is assessed as being either objective or subjective, and, within the subjective category either positive or negative. As this scheme is applied to synsets, it addresses the issue that the same word may have different senses that have different opinion-related properties.

The sum of the three scores is always 1.0. This allows the score to reflect the fact that a synset may have opinion-related properties to a certain degree. An example given by Esuli and Sebastiani is the synset [estimable (3)] with the sense “may be computed or estimated”. This has an Obj score of 1.0, and Pos and Neg scores of 0.0. This indicates that the sense is clearly classified as non-subjective. In contrast, the synset [estimable (1)] with the sense (gloss) “deserving of respect or high regard” has a Pos score of 0.75, a Neg score of 0.0 and an Obj score of 0.25.

In effect, this method allows senses to be classified as having multiple aspects to varying degrees. This idea was originally introduced by Kim and Hovy [8], but Esuli and Sebastiani’s SWN gives it functionality. They argue that the application of tools that grade opinion-related properties will play an important role in the future of text mining.

2.2 SentiWordNet 3.0

SWN 3.0 is a development of SWN 1.0 [5]. The intermediate development stages SWN 1.1 [11] and SWN 2.0 [12] were primarily for internal use of the developers and it is standard to compare SWN 3.0 directly with its public predecessor SWN 1.0. There are three primary differences between SWN versions 1.0 and 3.0:

a) Version 3 was developed as an annotation of WN 3.0 whereas version 1 applied to WN 2.0. This amounts to an updating of the resource, and has implications when it comes to comparing the accuracy of version 3.0 with version 1.0.

b) The main difference and the most important one for improving the accuracy of the resource is that an additional analytical stage is added after the semi-supervised learning stage associated with SWN 1.0. This is an iterative random-walk step that is carried out on the result of the semi-supervised learning algorithm.

c) In SWN 1.0, the glosses from WN 2.0 are used in the training stage, not the synsets themselves, with the result that the classifier is a gloss classifier rather than a synset classifier. In SWN 1.0, the gloss is a non-disambiguated collection of words. Esuli and Sebastiani call this a "bag of words" approach, where no attempt is made to determine the sense of the words (to disambiguate them). They are simply given a frequency-weighted score. This is different from SWN 3.0, where the development of the random-walk step requires the gloss to be disambiguated: in other words, to yield a further collection of synsets. Thus, in contrast to SWN 1.0, SWN 3.0 construction uses a "bag of synsets" approach.

A. Random-Walk step: The Concept

This step makes an assumption about relationships between words. The definiens (the words defining the word in question) is in a binary relationship with the definiendum (the word being defined). The assumption that underlies the random-walk step is that a direct link can be posited between synset 1 (S1) and synset 2 (S2), if and only if S1 occurs in the definiens of synset 2. The idea underlying this assumption is that if many words in the definiens are positive (or negative) then it is plausible that the definiendum is positive (or negative) too. This assumption allows positivity (or negativity) to, in the words of Baccianella et al. [5], "flow through the graph from the words used in the definitions to the words being defined."

B. Difficulty in implementing the random-walk step

A major issue is that, in WN, although the definiendum is a synset, the words used in the glosses are not. That is, the definiendum is disambiguated (different senses of the word are separated out), whereas the words in the definiens are non-disambiguated (the sense in which the word is being used is not made clear). This is not suitable for the random-walk step, which implies links between the words in both definiens and definiendum. For this to work, the glosses themselves must consist of a string of synsets (disambiguated words).
In the case of SWN 3.0, the Princeton WordNet Gloss Corpus[2] was used to achieve this. This contains manually disambiguated glosses for WN 3.0.

C. Assessment of SentiWordNet 3.0 Including the Random-Walk step, in relation to SentiWordNet 1.0

Baccianella et al. assessed whether the addition of the random-walk step improves or reduces the accuracy of the results delivered by SWN 3.0 in comparison with SWN 1.0, which lacked this element. Both SWNs were assessed using a small manually annotated subset of WN which was then compared with the automatic annotations of the same synsets by the respective versions of SWN.

The methodology used in their assessment is described below. There are two main difficulties that the methodology had to address:

a) How to create an "objective" criterion or "gold standard" against which the results of the SWN can be compared.

b) How to make an assessment that has validity across different versions of WN.

SWN 1.0 was evaluated on Micro-WN (Op) [16] which consists of 1,105 synsets of WN 2.0 that were manually annotated for sentiment by 5 people. The methodology involved the 5 annotators

[2] http://wordnet.princeton.edu/glosstag.shtml

working together to develop a common assessment understanding, and then working individually to increase speed. All synsets were rated by more than one annotator and results averaged. As with SWN 1.0 itself, they scored for Pos, Neg and Obj with a rating that added up to 1.0. The manual results can then be compared with SWN 1.0.

The difficulty with applying the same method to an assessment of SWN 3.0 is that SWN 3.0 applies to WN 3.0, which has different synsets from WN 2.0. To make a meaningful assessment of relative accuracy, the synsets that were manually assessed in WN 2.0 must first be mapped across to WN 3.0.

Baccianella et al. acknowledge the limitations of this process. They used three tools that were applied consecutively, with a later tool only being used where an earlier one failed to produce a result. The three tools are:

a) WN sense mappings (nouns and verbs only).

b) Synset word matching (if a synset contains the same words in Micro-WN(Op) and WN 3.0, and uniquely so, then they are considered to describe the same concept).

c) Gloss similarity (the greatest similarity between glosses determines the most likely equivalence of sense). They examined some results manually and found them to be satisfactory, but recognise that the results of the mapping process have not been completely checked for correctness: this would imply a complete search of WN 3.0 synsets of the same polarity to find the best match for each word in Micro-WN(Op).

D. Issues concerning SentiWordNet 3.0

The matching process allows the results of SWN 1.0 and SWN 3.0 to be compared. The rankings for Positivity and Negativity between SWM 1.0 and SWN 3.0 that were in complete agreement with the manual ranking would have a value of 0, complete disagreement a value of 1. Thus, the lower the value, the greater the agreement.

SWN 3.0 appears significantly more reliable than version 1.0 with a 19.48% increase in ranking by positivity and 21.96% by negativity. Even given the qualifications that were discussed above, this appears to be a marked improvement.

Important issues are raised by the development of SWN 3.0:

a) It is difficult to establish a gold standard against which to measure the effectiveness of an opinion mining resource. This becomes an even greater problem as the development of the target resource (WN) changes. Comparisons are difficult and likely to become more so.

b) SWN 3.0 has evolved by an additive process, adding steps. There is room for debate whether future development will be best achieved by continuing the same process — adding steps that refine the results further — or by revisiting and refining some existing steps.

4. GENERATING ARABIC SENTIWORDNET

Arabic is a widely used language that has both economic and political importance. It is natural that tools that enable sentiments in Arabic texts to be extracted and assessed are of great interest. Existing development work has been focused on developing for Arabic the equivalent of the resources that exist for English. By adding sentiment information to the Arabic WN to generate the Arabic SWN, we have improved the set of WN-based resources for the benefit of researchers in Arabic Natural Language Processing (NLP). This has the potential to enable future opinion mining tools to be developed directly for Arabic texts.

The key initial requirement for the application of SWN to Arabic is a version of WN on which it can operate. There is an existing Arabic version of WN 2.0 [13, 14] — Arabic WN 2.0[3], but not of WN 3.0.

Therefore, we built the database for Arabic SWN taking into account all the preparation levels shown in Figure 1:

a) The Arabic WN 2.0 database was upgraded to version 3.0 by mapping to the latest English WN 3.0 database.

b) The English SWN 3.0 database was also mapped to the new version of the Arabic WN 3.0 database.

c) All fields in our new database — Arabic WN 3.0 — were checked and revised with the English SWN 3.0, and then only the fields that existed in the English version (which express sentiments) were kept in the Arabic version to give us the Arabic SWN, with the rest deleted.

Through this mapping, an Arabic SWN database was generated. However, it contains fewer words and there is still a need to translate all the other fields that exist in the English version. Thus, the total number of words existing in the Arabic SWN is around 10,000, which includes verbs, adjectives, adverbs and nouns.

3.1 Approaches for using the Arabic SWN

Two different approaches can be used for developing the Arabic SWN:

a) Using the database we built as a multi-lingual setup to be applied to both English and Arabic contexts.

b) Having all synsets in the English version translated into the Arabic version – which is the approach we used in our work.

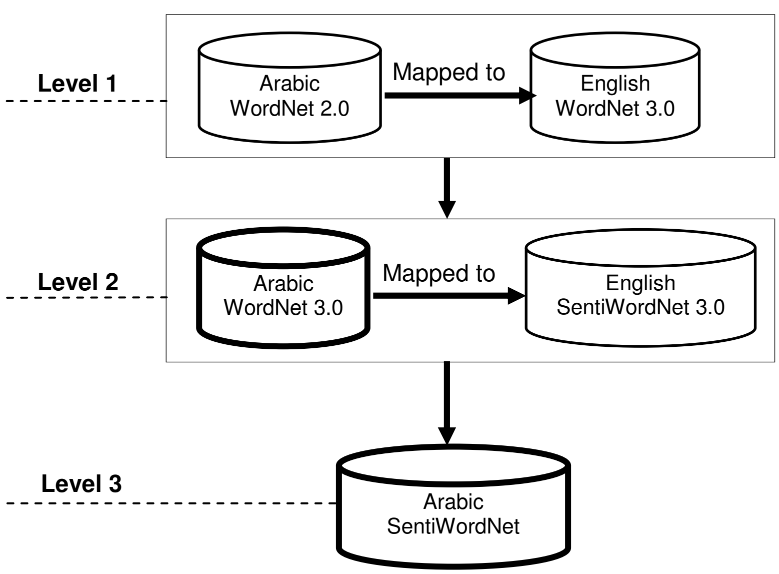

FIGURE 1: Preparation step for generating Arabic SWN.

[3] http://www.globalwordnet.org/AWN/

3.2 The AWN-WordNet Mappings

When the AWN was constructed, each of its synsets was mapped to Princeton WordNet (for English) Version 2.0 despite two more recent versions of Princeton WordNet being released while the AWN was under construction. However, SentiWordNet provides additional annotations on WordNet version 3.0, so it was necessary to update the Arabic to English mapping to WordNet 3.0. This was done using the mapping from WordNet 2.0 to 3.0 available from http://nlp.lsi.upc.edu, which was constructed automatically using the procedure described in [30]. The mappings give a unique WN3.0 synset ID for 99.7% of WN2.0 adverb synsets, 98.77% of adjective synsets, 99.39% of noun synsets and 98.92% of verb synsets. Of the ambiguously mapped WN2.0 synsets, a majority have no AWN linkage, but some remained to be verified manually.

3.3 Methodology

Several research studies have been done on opinion mining and sentiment analysis using reviews and movies as their datasets, either in English or Arabic [20, 27, 28, 29]. For the purpose of our research, to generate a new corpus for Arabic sentiment analysis, the data was generated from Arabic social media in the form of technology blogs (2,350 sentences). Technology blogs provide various challenges:

a) Use of a foreign language (English in our case) for names of technologies, companies, software or programs.

b) Transliterations within the texts.

c) Difficulty of recognizing opinions and sentiments expressed to show users' opinions towards these technologies and companies.

One of the purposes of building this corpus is to evaluate the sentiment coverage of the Arabic SWN. Our corpus construction methodology involves the manual detection of sentiments via manual annotation, which was carried out according to our annotation guidelines, which were:

a) Determine positive or negative words about companies.

b) Determine positive or negative words about technologies, products or systems.

c) For each sentence, each sentiment word should be detected as positive or negative and then determination is made whether the whole sentence is positive or negative.

Finally, the annotated sentiment words are compared with the words that exist in the Arabic SWN for the evaluation step (section 3.4). Tables 3 and 4 show some Arabic examples of the sentiment annotation tasks about companies and technologies, products or systems, respectively (note: all examples are translated into English for ease of understanding):

Companies	
A positive sentence	***A negative sentence***
خطوة كبيرة من الفيس بوك بتوفير هذا المتجر والذي سينجح بكل سهولة.	كاسبرسكي: أبل تفتقد للجدية في حماية نظام الماك
A big step from Facebook to provide this store, which will succeed easily.	Kaspersky: Apple is not serious about protecting the Mac system.

TABLE 3: Examples of the sentiment annotation tasks about companies.

Technologies, products or systems	
A positive sentence	**A negative sentence**
أحب استخدام خدمة فورسكوير لتحديد موقعي.	شخصياً أرى أن مستقبل نظام الاندرويد بدأ يصبح مخيفاً وغير آمناً.
I love using foursquare to determine my location.	Personally, I think the future of Android is becoming scary and unsecure.

TABLE 4: examples of the sentiment annotation tasks about technologies, products or systems

3.4 Discussion of Annotation Results

It was found that 31% of sentences were annotated with negative sentiment and 64% with positive. The inter-annotator agreement (IAA) is shown in figure 2.

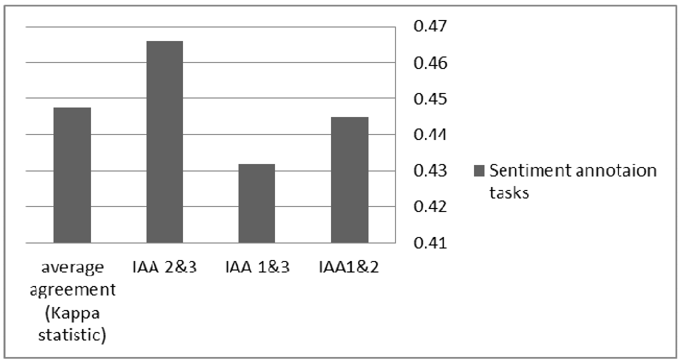

FIGURE 2: The Inter-annotator agreement (IAA) and average agreement (Kappa statistic)

Three annotators were involved and the Kappa statistic was used to calculate IAA for each sentence: the overall result obtained was 0.45 which is considered as "moderate agreement" [17].

We intend to improve on this result by adopting a cyclic annotation approach. However, we must first establish the sentiment coverage of the Arabic SWN. We therefore carried out an evaluation step to check which annotated sentiment words occur in the Arabic SWN. We already know that the Arabic SWN includes fewer synsets than English SWN. It was established that around 5% of annotated sentiment words did not occur in the Arabic SWN. This result is clearly influenced by the size and nature of our corpus. However, taken together with the difference in the number of synsets between Arabic SWN and the English SWN, it is evident that we must expand the coverage of Arabic SWN. Returning to our cyclic annotation approach we plan to speed up and improve annotation by incorporating automatic look up of Arabic SWN to tag recognised sentiment words for validation by the annotators, which will reduce the annotation burden. We will though firstly translate the missing synsets in Arabic SWN from English SWN. As this is a translation exercise it is to be expected that there will be Arabic specific sentiment synsets that will be still missing. Some of these may be added from our corpus, once we are able to determine the complete coverage of the Arabic SWN augmented by the translated synsets.

5. Related Work

Due to the lack of an existing SentiWordNet lexicon for the Arabic language, several research studies on opinion mining and sentiment analysis for Arabic have used lists of sentiment words to cover their research needs [18, 19, 20]. Recently, a study done by Abdul-Mageed and Diab [21] focussed on expanding a polarity lexicon of Modern Standard Arabic built manually by leveraging various existing polarity lexica for English. The utility of their expanded lexicon was not tested and moreover they focussed on adjectives.

Unfortunately, a comparison cannot be made between our Arabic SWN and previous systems because they have not yet been made available; furthermore, our database links to the original Arabic WordNet which is not the case for other previous systems.

6. CONCLUSION AND FUTURE DIRECTIONS

SWN has proved to be an effective tool for opinion mining and sentiment classification, and the improvement in effectiveness that has been achieved between versions 1.0 and 3.0 gives grounds for optimism that it provides a tool with further developmental possibilities. That is, it is flexible enough to accommodate both additional refining steps and new work on the existing steps.

In this paper, we reported efforts to generate an Arabic SWN database in relation to the English SWN 3.0. There is however a need to increase the number of words in our Arabic SWN. We plan also to refine further the annotation of our corpus to achieve a higher IAA score, especially as we plan to use this corpus for research purposes in Arabic sentiment analysis. The ranking process for positivity and negativity is another step to be taken in the near future. All these limitations above need to be resolved in future to enhance the final version of our Arabic SWN which will be made publicly available.

Acknowledgments

This research was supported via a PhD scholarship from Saudi Cultural Bureau.

7. REFERENCES

1. A. Esuli, and F. Sebastiani, “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining.” in Proc. of the 5th Conference on International Language Resources and Evaluation (LREC’06), 2006, pp. 417–422.

2. B. Pang, and L. Lee, “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.” in Proc. of the ACL, 2004, pp. 271–278.

3. B. Liu, “Opinion Mining and Sentiment Analysis.” in B. Liu, Web Data Mining, Springer Berlin, Heidelberg, 2011, pp. 459–526.

4. T. Wilson, J. Wiebe, and R. Hwa, “Just how mad are you? Finding strong and weak opinion clauses.” in Proc. of AAAI, 2004, pp. 761–769.

5. S. Baccianella, A. Esuli, and F. Sebastiani, “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.” in Proc. of the 7th Conference on International Language Resources and Evaluation (LREC’10), 2010, pp. 2200–2204.

6. H. Takamura, T. Inui, and M. Okumura, “Extracting Emotional Polarity of Words using Spin Model.” in Proc. of ACL-05, 43rd Annual Meeting of the Association for Computational Linguistics, 2005, pp. 133–140.

7. A. Esuli and F. Sebastiani, “Determining the semantic orientation of words through gloss classification.” in Proc. of the 14th ACM international conference on Information and Knowledge Management ACM, 2005, pp. 617–624.

8. S.-M. Kim and E. Hovy, “Determining the sentiment of opinions.” in Proc. of the 20th international conference on Computational Linguistics Association for Computational Linguistics, Geneva, Switzerland, 2004, pp. 1-1367.

9. A. Esuli, and F. Sebastiani, “Determining Term Subjectivity and Term Orientation for Opinion Mining.” in Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL’06), 2006, 193-200.

10. A. Valitutti, C. Strapparava and O. Stock, “Developing Affective Lexical Resources.” Psychology journal, 2004, pp. 61–83.

11. A. Esuli and F. Sebastiani, “SentiWordNet: a high-coverage lexical resource for opinion mining.” Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR), 2006.

12. A. Esuli, “Automatic generation of lexical resources for opinion mining: models, algorithms and applications.” SIGIR Forum, ACM, vol. 42(2), 2008, pp. 105–106.

13. W. Black, S. Elkateb, and P. Vossen, “Introducing the Arabic WordNet Project.” in Proc. of the third International WordNet Conference (GWC-06), 2006, 295-300.

14. S. Elkateb, W. Black, D. Farwell, P. Vossen, A. Pease, and C. Fellbaum, “Arabic WordNet and the Challenges of Arabic.” In Proc. of the British Computer Society Conference on Challenges of Arabic for Natural Language Processing and Machine Translation.The British Computer Society (BCS), London, 2006, pp. 15–25.

15. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller. "Introduction to wordnet: An on-line lexical database*." International journal of lexicography, vol. 3(4), 1990, pp. 235-244.

16. S. Cerini, V. Compagnoni, A. Demontis, M. Formentelli, and G. Gandini. "Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining," Language resources and linguistic theory: Typology, second language acquisition, English linguistics, A. Sans`o, Ed. Milano: IT, 2007, pp. 200–210 .English linguistics, pp. 200–210. Franco Angeli Editore, Milano, IT.

17. A. J. Viera and J. M. Garrett. "Understanding interobserver agreement: the kappa statistic." Family Medicine, vol. 37, 2005, pp. 360–363.

18. A. A. Al-Subaihin, H. S. Al-Khalifa and Al-Salman, A. S. "A proposed sentiment analysis tool for modern Arabic using human-based computing." In Proc. of the 13th International Conference on Information Integration and Web-based Applications and Services, Ho Chi Minh City, Vietnam, ACM**,** 2011, pp. 543-546.

19. N. Farra, E. Challita, R. Abou Assi and H. Hajj. "Sentence-Level and Document-Level Sentiment Mining for Arabic Texts." in Proc. of the 2010 IEEE International Conference on Data Mining Workshops, IEEE Computer Society, 2010, pp. 1114-1119.

20. M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López and J. M. Perea-Ortega. "OCA: Opinion corpus for Arabic." J. Am. Soc. Inf. Sci. Technol, vol. 62, 2011, pp. 2045-2054.

21. M. Abdul-Mageed and M. Diab. "Toward Building a Large-Scale Arabic Sentiment Lexicon." in Proc. of the 6th International Global WordNet, 2012, pp. 9-13.

22. B. Liu, X. Li, W. S. Lee and P. S. Yu. "Text classification by labeling words." in Proc. of the National Conference on Artificial Intelligence, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press, 1999, pp. 425-430.

23. X. Ding and B. Liu."The utility of linguistic rules in opinion mining." in Proc. of the 30th annual international ACM SIGIR conference on Research and development in information retrieva, 2007, pp. 811-812.

24. T. Wilson, J. Wiebe and P. Hoffmann. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis." in Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, Association for Computational Linguistics, 2005, pp. 347-354.

25. P. J. Stone, C. D. Dexter, S. S. Marshall and M. O. Daniel. "The General Inquirer: A Computer Approach to Content Analysis." Cambridge, MA: MIT Press, 1966.

26. J. W. Pennebaker,R. J. Booth and M. E. Francis. "Linguistic inquiry and word count: LIWC 2001." Mahway: Lawrence Erlbaum Associates, 2001.

27. R. Prabowo and M. Thelwall. "Sentiment analysis: A combined approach." J. Informetrics, vol. 3, 2009, pp. 143-157.

28. P. D. Turney. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." in Proc. of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania, Association for Computational Linguistics, 2002, pp. 417-424.

29. M. Hu and B. Liu. "Mining and summarizing customer reviews." in Proc. of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. Seattle, WA, USA, ACM, 2004, pp. 168-177.

30. J. Daudé, L. Daudé and G. Rigau. " Mapping WordNets Using Structural Information." in 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000), Hong Kong, 2000.

"C'mon – You Should Read This": Automatic Identification of Tone from Language Text

Lisa Pearl *lpearl@uci.edu*
Department of Cognitive Sciences
University of California, Irvine
Irvine, CA 92697, USA

Mark Steyvers *mark.steyvers@uci.edu*
Department of Cognitive Sciences
University of California, Irvine
Irvine, CA 92697, USA

Abstract

Information extraction researchers have recently recognized that more subtle information beyond the basic semantic content of a message can be communicated via linguistic features in text, such as sentiments, emotions, perspectives, and intentions. One way to describe this information is that it represents something about the generator's mental state, which is often interpreted as the *tone* of the message. A current technical barrier to developing a general-purpose tone identification system is the lack of reliable training data, with messages annotated with the message tone. We first describe a method for creating the necessary annotated data using human-based computation, based on interactive games between humans trying to generate and interpret messages conveying different tones. This draws on the use of *game with a purpose* methods from computer science and *wisdom of the crowds* methods from cognitive science. We then demonstrate the utility of this kind of database and the advantage of human-based computation by examining the performance of two machine learning classifiers trained on the database, each of which uses only shallow linguistic features. Though we already find near-human levels of performance with one classifier, we also suggest more sophisticated linguistic features and alternate implementations for the database that may improve tone identification results further.

Keywords: Language Text, Mental States, Tone, Game with a Purpose, Information Extraction.

1. INTRODUCTION

One focus of information extraction research has been identifying basic semantic content (e.g., identifying who did what to whom when). Recently, however, researchers have recognized that more subtle information can be communicated via linguistic features in text (see [1] for a review), and this has spurred research in sentiment analysis [2][3][4][5][6][7][8][9], emotion and speech act identification [10][11][12], perspective identification [13][14], and deception detection [15][16][17][18][19] in language text. All of these research areas have in common the basic idea that humans have a mental state that they express in the messages they create, in addition to the basic semantic content of those messages. This mental state can be an emotion like anger or embarrassment, an attitude like confidence or disbelief, or an intention like persuasion or deception (among other things), and it is often perceived as the *tone* of the message. For example, in "C'mon – you should read this", the basic semantic content is something like *read(you, this)* while the tone of the message is persuasive. A message's tone is instrumental in understanding the underlying mental state (and motives) of the person who generated the message, and for predicting how this message will be interpreted by humans reading it.

While most text software is equipped with a spell checker and a grammar checker, no programs currently offer a "tone checker" – though this would be a highly desirable feature to have. As one example, it could be used to check that an email will not be interpreted in a way the sender didn't intend, such as offensive, pushy, or stilted. One technical barrier to developing a general-purpose system dedicated to tone identification in text is a lack of reliable data on which to train it. Currently, there are no large-scale text databases annotated with tone information, and so it is difficult to know what linguistic features to use as cues. Specifically, there are few examples of the "ground truth" for tone, i.e., text annotated with human perceptions of the message's tone. Given the success of natural language processing methods in the other information extraction endeavors mentioned above, we believe tone could also be retrievable once the relevant linguistic cues are identified.

One way to create the necessary annotated data is to make use of human-based computation [20], since humans are used to transmitting messages with specific tones and interpreting the tone of messages. We first describe a methodology for creating this kind of database that we have implemented in the form of a *game with a purpose* (GWAP) [21][22][23], and discuss some salient properties of the resulting database. We then present results from two machine learning classifiers trained on data from the database, which demonstrate the utility of the database, the advantage of human-based computation, and the benefits and pitfalls of using shallow linguistic features for identifying tone. We conclude with suggestions for more sophisticated linguistic features based on the current results, as well as some discussion of what constitutes the "ground truth" for tone information.

2. A RELIABLE DATABASE FOR TONE

2.1 The need for databases

In general, reliable databases are required to develop reliable machine learning algorithms. Unfortunately, few databases annotated with mental state information exist, and these are generally small in size compared to corpora generally available for natural language processing (e.g., the English Gigaword corpus [24] contains approximately 1.75 million words; the Personal Story Subset of the Spinn3r Blog Dataset [25][26] contains approximately 5.3 million words). A few recent examples demonstrate this.

The Language Understanding Annotation Corpus (LUAC) [27] includes text annotated with committed belief, which "distinguishes between statements which assert belief or opinion, those which contain speculation, and statements which convey fact or otherwise do not convey belief." This database is meant to aid in determining which beliefs can be ascribed to a communicator and how strongly the communicator holds those beliefs. The LUAC contains only about 9000 words across two languages (6,949 English, 2,183 Arabic).

The Bitter Lemon corpus [14] is a compilation of essays on various Middle East issues, written from both Israeli and Palestinian perspectives. It is derived from a website (http://www.bitterlemons.net) that invites weekly discussions on a topic and publishes essays from two sets of authors each week. It has been used to investigate automatic classification of perspective [13][14], and contains 297 essays averaging 700-800 words each, for a total of approximately 22,000 words.

A corpus of blog posts annotated for persuasive tactics such as moral generalizations and redefinition was compiled by Anand and colleagues [11], for the purpose of training machine learning algorithms to recognize when persuasion is intended. This corpus contains 380 blog posts (as the rest of the approximately 25,000 examined did not contain easily identifiable persuasive acts). Based on estimates from the larger blog corpus from which these posts were taken [28], there were approximately 200 words per post, leading to a corpus size of approximately 76,000 words.

The last corpus demonstrates an additional issue that surfaces when trying to find realistic examples of language expressing different mental states: It may well be that most of the available data does not actually express any of the mental states of interest. One way around this is to look for open-source data that are highly likely to express the mental states of interest by happenstance, e.g., online gaming forums with games that happen to involve deception (e.g., Mafia game forums [19]). This can often lead to datasets that may be larger in size. However, another issue remains, even in these situations where larger quantities of open-source data are available: The breadth of coverage is limited. While it may be easy to find examples of some mental states in open-source data, it is not always easy to find all the ones of interest.

Moreover, real world data sets present the basic problem of ground truth, i.e., knowing for certain which mental states were intended to be conveyed by a particular message. Human annotators can attempt to recover this information, which is the approach taken by the annotated corpora mentioned above. However, this is often a time-intensive and human-resource-intensive process.

2.2 Using games with a purpose

Notably, the human annotation process used for previous corpora highlights that humans are in fact able to interpret the mental state behind a message. Human-based computation can leverage this ability from the population, and use it to construct a reliable database of messages expressing different mental states. Interestingly, groups of humans are sometimes capable of producing more precise and reliable results than any particular individual in the group. This kind of "wisdom of the crowds" phenomenon has been demonstrated in many knowledge domains, including human memory, problem solving, and prediction [29][30][31][32]. Snow and colleagues [33] have additionally demonstrated that a relatively small number of non-expert annotations in natural language tasks can achieve the same results as expert annotation. This suggests that this approach on a larger scale will likely be able to yield a reliable annotation of mental state information in language, as expressed by message tone.

One approach is to use a game with a purpose (GWAP) [21] that is designed to encourage people to provide both kinds of data needed in the database: the messages expressing a particular tone and the annotation of that tone for every message. GWAPs are currently used to accumulate information about many things that humans find easy to identify (see http://www.gwap.com/gwap/ for several examples), such as objects in images [22], common sense relationships between concepts [23], belief about others' preferences [34], and the musical style of songs [35]. Using a GWAP for message tone, we can also take advantage of human-based computation and the potential wisdom of the crowds inherent in this setup. In particular, since the collected data come from and are vetted by a large number of participants, we can gauge which messages are reliable examples of particular mental states and which are confusing examples.

We have designed a GWAP called Word Sleuth (available at http://gwap.ss.uci.edu/) that takes the form of an online game played through a web browser interface. In the context of the game, Word Sleuth encourages participants to play two different roles, which participants can freely alternate between as they desire: the message generator (called the Expressor) and the message interpreter (called the Word Sleuth). As the Expressor, participants generate messages expressing a particular tone; as the Word Sleuth, they label messages created by other participants as expressing a particular tone. Game play is asynchronous, so participants do not need a partner to play. Figures 1 and 2 show example game play for both the Expressor and Word Sleuth roles.

FIGURE 1: An Example of Expressor Game Play.

The first panel of figure 1 shows a sample screen for generating a persuasive message. The participant is shown a random context picture to help them generate a message, and this context picture will be shown to any interpreter attempting to interpret the generated message. Participants can choose what difficulty level they want to play, with harder difficulty levels earning them more points. The difficulty level for Expressors determines how many "taboo" words there are – these are words that the participant cannot use in the message. For the easiest level, this includes only morphological variants of the intended tone (e.g., "persuading", "persuasion", "persuades", and "persuaded" for the "persuading" tone). Harder levels add additional words (3 for medium difficulty, 7 for hard difficulty) that are dynamically generated based on the messages that have previously been created for that tone – words that currently have the highest mutual information are included in the taboo list. The motivation behind taboo words is two-fold. First, the morphological variants discourage certain kinds of cheating (e.g., explicitly writing "This is a persuading message"). Second, the dynamically generated taboo words encourage participants to create more varied messages, rather than relying on a few key words.

Participants are additionally shown the other potential message tones that interpreters will have to choose from. This encourages them to make their messages unambiguously express the intended tone. Participants are also reminded that there is a mechanism for flagging poor messages – if a generator's message is flagged as poor by enough other participants, that message is removed from game play and the generator's expressive score is penalized. Because the intended tone type is randomly selected, it is possible to get the same type multiple times in a row when generating messages. Given this, we also give participants the option to skip creating a message – this is useful when they don't wish to generate a message for a particular tone (perhaps because they just generated a message of that kind).

The second panel of figure 1 displays a sample follow-up for message creation. The participants can see that their activity points have increased, and are encouraged to check in later to see if their expressive score has increased. They then have the option to continue the current game mode (e.g, Expressor, hard difficulty), or change either the role or difficulty or both.

The first panel of figure 2 shows a sample screen for interpreting a message's tone. The message, which was generated by another participant previously, is shown, along with the context picture that the generator used to create it. The difficulty level is shown, with higher difficulty levels gaining the participant more points if the message's tone is interpreted correctly. A message's difficulty is determined by how accurate other people have been so far at interpreting its intended tone – the third of the message set with the lowest accuracy is labeled "hard", the middle third is labeled "medium", and the top third is labeled "easy". The second panel of figure 2 shows a sample screen that would appear after a message's tone has been chosen from among the available options. The intended tone is displayed along with the interpreted tone. If they match, the receptive score is shown to increase and the receptive IQ may increase. If they do not match, the receptive score is shown to decrease and the receptive IQ may decrease. In either event, the activity points increase (with the aim of motivating continued game play). Participants also have an option at this point to flag a message if they believe it is a particularly poor example of a message expressing the intended tone. This helps to immediately weed out very poor examples from the database.

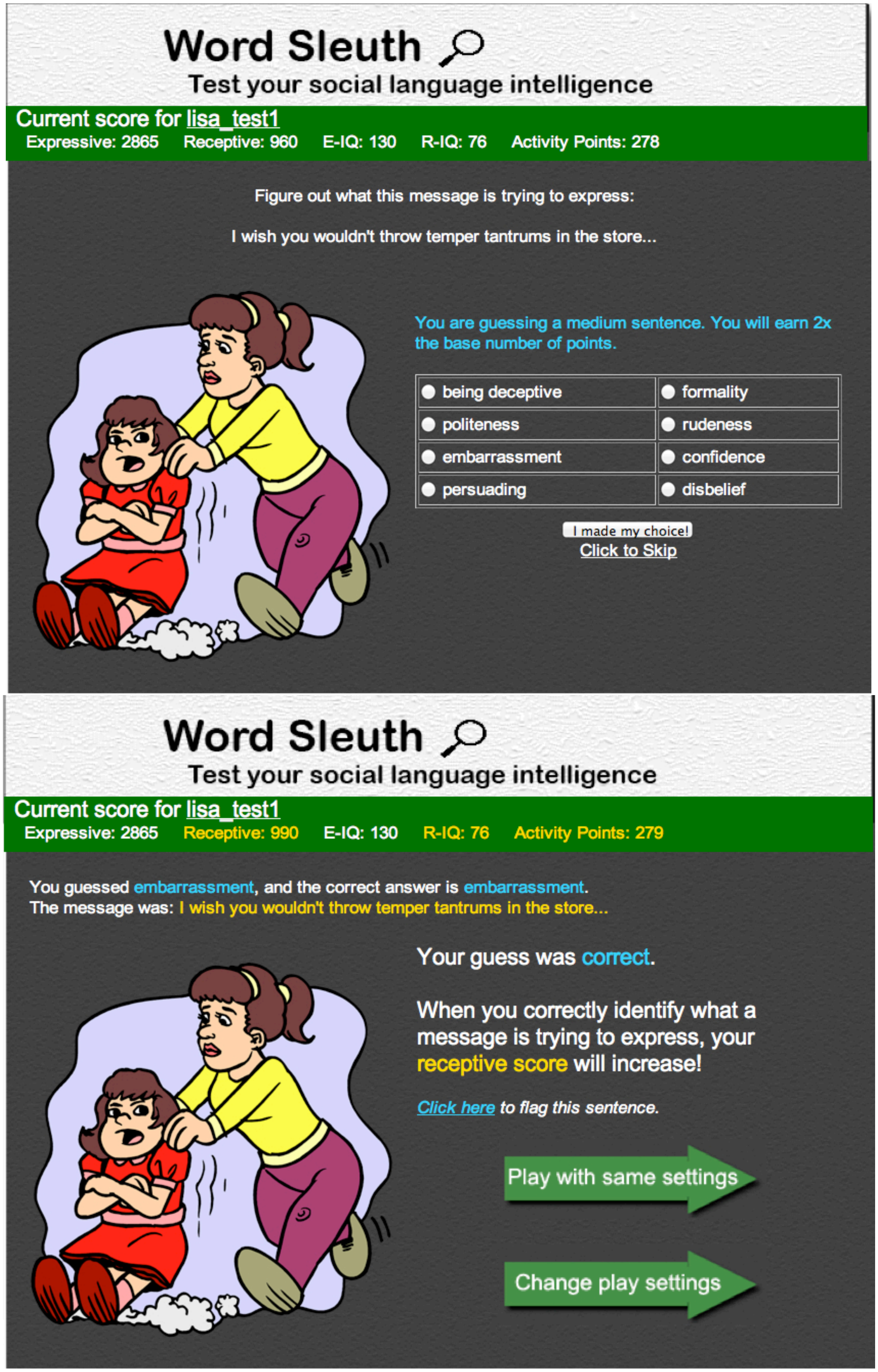

FIGURE 2: An Example of Word Sleuth Game Play.

The scoring system within the game reflects the inherent connection between the generator's ability, the message, and the interpreter's ability. When a message's tone is correctly interpreted, both the generator of that message and the interpreter of that message get points added to their scores – the generator's expressive score increases while the interpreter's receptive score increases. In addition, the generator's "expressive IQ" increases and the interpreter's "receptive IQ" increases, based on z-scores of overall percent correct expressing or interpreting, respectively. When a message is not labeled correctly (whether due to the generator creating a poor message or the interpreter interpreting it poorly), no points are subtracted from the respective expressive and receptive scores – however, the expressive and receptive IQs are decreased (again, based on z-scores of percent correct expressing or interpreting). Because scores and IQs are updated only when messages are interpreted, we additionally have activity points which are increased whenever a participant either creates a message or guesses the label for a message. This is meant to particularly encourage participants to play the Expressor role, as activity points give them instant gratification for creating a message. To encourage game play in general, there are high score tables available, as well as individual achievement badges.

With enough game players, many messages expressing different tones can be created and interpreted. Given previous successes with human computation and wisdom of the crowds effects, we expect the cumulative knowledge to be quite reliable, even if a message is only labeled with a single tone (perhaps expressing that message's most obvious tone from the perspective of the interpreter). This is because the same text can be evaluated by many different people, which can reduce the effect of idiosyncratic responses from a few individuals.

One clear advantage of the GWAP approach is the ability to target which mental states we are interested in, and subsequently create messages with those tones that are also annotated with that tone information. In this way, we simultaneously address three issues that previous databases have encountered. First, we know that nearly all our data[1] will in fact contain at least one of the tones of interest, because the data have been generated to express those very tones. Second, we can infer tone annotations in a fairly inexpensive way – using the collective interpretations of the Word Sleuth game players. Third, we can be fairly sure of the accuracy of the annotation by using a wisdom of the crowds approach that aggregates data across multiple interpretations.

2.3 Creating a tone database with Word Sleuth

We decided to explore eight mental states that are indicators for different emotional states, attitudes, and intentions: deception, politeness, rudeness, embarrassment, confidence, disbelief, formality, and persuading. As of March 2012, Word Sleuth has attracted 877 online game players, with 4,157 messages generated (~48,500 words) and 29,586 interpretations of those messages (an average of about 7 interpretations per message). Participants generally played the Word Sleuth role more than the Expressor role, which has led to significantly more interpretations as compared to messages. There was no limit on message length, though more participants tended to keep messages fairly brief (approximately 11-12 words).

Averaged across messages and participants, humans were successful at mental state transmission (via message tone) approximately 74.4% of the time. That is, approximately 3 out of every 4 messages were interpreted as expressing the tone they were intended to express. This is significantly better than chance performance, which would be 1 out of 8 (12.5%), and demonstrates that humans are fairly good at transmitting message tone – though notably not perfect. Moreover, human accuracy is not evenly distributed across the different tone types, as

[1] The messages flagged during game play as being very poor are the only data that might not be useful.

shown in figure 3. Figure 3 is a confusion matrix that shows the likelihood that a message will be interpreted as a specific tone (in the columns), given that it has been generated with that specific tone in mind (in the rows), averaged over messages and participants. In other words, figure 3 shows the conditional probability distribution *p(interpreted | generated)*. The diagonal probabilities indicate how often a message's tone was correctly interpreted for each tone type; this shows how often transmission of that particular mental state was successful. The total number of interpretations for each tone type is shown in the rightmost column. While the messages are chosen randomly for interpretation when participants play the Word Sleuth role, participants do have the option to skip messages they find difficult – this is what causes certain tone types, such as deception and formality, to be less represented in the dataset. In essence, this is one indication of the inherent difficulty of those two tone types.

	deception	**politeness**	**rudeness**	**embarrassment**	**confidence**	**disbelief**	**formality**	**persuading**	**total interpretations**
deception	**0.59**	0.05	0.05	0.03	0.07	0.05	0.02	0.14	3178
politeness	0.02	**0.72**	0.02	0.02	0.02	0.01	0.13	0.07	3435
rudeness	0.01	0.01	**0.86**	0.02	0.02	0.04	0.01	0.03	4320
embarrassment	0.03	0.05	0.02	**0.77**	0.01	0.08	0.02	0.02	3936
confidence	0.03	0.03	0.02	0.01	**0.81**	0.02	0.01	0.07	4174
disbelief	0.03	0.03	0.05	0.03	0.03	**0.80**	0.01	0.02	4415
formality	0.02	0.35	0.02	0.02	0.04	0.02	**0.46**	0.08	2314
persuading	0.05	0.05	0.02	0.00	0.07	0.01	0.02	**0.77**	3814
									29586

FIGURE 3: Human confusion matrix for the eight tone types investigated. The rows represent the intended tone, while the columns represent the interpreted tones. The bolded diagonal indicates the percentage of correct interpretations for each tone type. The total number of interpretations for each tone type is shown in the rightmost column.

Another indication is the accuracy of transmission – deception (0.59) and formality (0.46) are much harder to transmit correctly than the other tone types. Table 1 shows some sample messages (with the participants' own spelling and punctuation), highlighting why some tone types may be easier than others.

Intended Tone	Interpreted Tone	Message
confidence	confidence	"here's the paper! i'm positive its really good this time"
rudeness	rudeness	"You are the stinkiest person I've ever met."
deception	persuading	"I recommend that you take one step forward. Don't worry, it's not dangerous."
formality	politeness	"may i take the road on the left please"

TABLE 1: Sample messages created in the Word Sleuth game. The top two message tones are correctly interpreted, while the bottom two message tones are not.

Confidence (0.81) and rudeness (0.86) have the highest rates of successful transmission, as shown in figure 3, and Table 1 shows how these messages may be very distinctive. For example, the confident message uses an indicator of certainty ('positive'), while the rude message uses the negative valence word 'stinkiest'. In contrast to these two tone types, deception is more difficult and often confused with persuading (0.14) and confidence (0.07). This is likely because the linguistic cues overlap – when attempting deception, it may often be while in the act of persuading, and the deceiver may be attempting to appear confident in order to be believed. This appears in the example message in Table 1 – the overall message is attempting to persuade the listener to move forward. Since it is intended as deception, the most likely part that is deceptive involves the speaker's assessment of how dangerous the situation is (presumably, it *is* in fact dangerous to take a step forward). This highlights one way in which deception may be a more complex intention – it effectively involves a semantic inversion, where the opposite of the semantic content is actually true, and the participant must detect that inversion. This in contrast to the other tone types, which can be viewed as an adjusted version of the underlying semantic content (e.g., a persuasive form of *read(you, this)* does not change this underlying semantic content). Interestingly, while participants have difficulty detecting all the deceptive messages (that is, having good recall), their precision is fairly good (0.59/Σ(deception column) = 0.76) – i.e., when they have decided something is deceptive, it usually is.

The confusion matrix in figure 3 also shows that formality is often confused with politeness (0.35) – that is, formal messages are mistaken as polite about one third of the time. The sample message in Table 1 shows how this might happen – while the 'may ' construction is often used to convey formality (as opposed to 'can I'), other linguistic cues may have conveyed politeness more strongly to the interpreter, such as the 'please' at the end of the utterance. Precision is somewhat better (0.46/Σ(formality column) = 0.68), again showing that the messages interpreted as formal often are indeed formal. Interestingly, figure 3 also shows us that politeness is more accurately transmitted overall (0.72) and not as often confused with formality (0.13). This suggests that, though these tones do overlap, formality may be viewed by the participants as a subset of politeness. In effect, it is easier to be polite without being formal, and it seems more difficult to be formal without also being polite. This has some intuitive appeal, as a formal tone may be viewed as polite, even if the content of the message is not very polite (e.g., a complaint).

From a natural language processing standpoint, these human confusion data are useful in two ways. First, they give us a goal to aim at – we would like to eventually do as well as humans (and perhaps even better, by avoiding the confusions humans stumble over). Second, these data suggest that two tones – deception and formality – are likely to be more difficult to automatically classify.

3. LEARNING TO IDENTIFY TONE AUTOMATICALLY

To demonstrate the utility of this kind of database for developing automatic systems for tone detection, we investigated the performance of two machine learning classifiers that were trained on portions of the current database. While we realize that there are many machine learning techniques that could be used, we decided to examine one very simple classifier and one more sophisticated classifier in order to demonstrate the utility of the kind of database we have constructed.

The goal of each classifier was the same as that of the humans playing the Word Sleuth role: select the intended tone from one of the eight choices. Though the database is still small when compared to the standard corpora used for developing natural language processing systems, we nonetheless find quite good performance when using the kind of linguistic features often used in previous studies of sentiment analysis, emotion identification, perspective identification, and deception detection. Performance is enhanced when we apply a simple wisdom of the crowds measure for selecting reliable messages to train on. These promising results suggests that larger databases constructed in a similar fashion may well lead to human-level classification performance and beyond.

More broadly, we are interested in the linguistic features that are useful for tone detection in text, and how machine learning algorithms compare to human performance. In particular, if we can identify the linguistic features humans are using, we may able to increase the performance of machines to human levels. This can also help us understand why humans make the mistakes they do (e.g., on deception and formality), so that software can be designed to recognize those mistakes. To this end, we use the classifier results to suggest more sophisticated linguistic features that may be more similar to the ones humans use.

3.1 Setting up the classification task

There are several considerations for any classifier approach: what data are used to train the classifier, what is a reasonable baseline to compare performance against, and what features does the classifier use to make its decision? We look at each of these in turn.

Given the relatively small size of the current database, there are two approaches regarding the data we use to train the classifiers. The first approach is to use a message regardless of how reliable it is, with the idea that the quantity of messages will make up for poor examples. An alternative approach is to only use a message if it is reliable, with the idea that better quality messages will make up for having fewer of them. The first approach will use all 4,157 messages currently available. For the second approach, we defined a simple measure that draws on the wisdom of the crowds: if a message has two or more interpretations and also has more than 50% agreement with the intended tone, it is included. This rules out messages with only one interpretation (since that doesn't represent a crowd's collective interpretation), and also messages where there was so much confusion that the intended tone was chosen 50% or less of the time. Applying this metric, we are left with a dataset of 1,862 messages (~21,750 words).

Turning now to reasonable assessments of baseline performance, we have two reasonable options. One is based on the task itself – given that there are 8 options, and the classifier must select one, there is a 1 in 8 chance of doing so correctly by chance (0.125). A slightly more informed baseline might be to always choose the tone type that is most frequent in the training data. For the complete dataset, this is the rudeness tone, which accounts for 570 of the 4,157 messages (0.137). For the filtered dataset, this is again the rudeness tone, which accounts for 321 of the 1,862 messages (0.172). Each dataset and its accompanying baselines are summarized in Table 2.

Dataset	Description	# Messages	baseline: 1 in 8	baseline: most frequent in training set
complete	all	4,157	0.125	0.137
filtered	2+ interpretations and >50% agreement	1,862	0.125	0.172

TABLE 2: Summary of two datasets that the classifiers are trained on, including the description of messages included in the dataset, the number of messages in the dataset, and two baseline performance measures.

The next question is which features the classifiers will use. As a first pass measure, we examined a number of fairly shallow linguistic features similar to what previous studies in sentiment analysis, emotion and speech act identification, perspective identification, and deception detection have used [7][8][11][13][14][15][16][17][18][19][36]. Table 3 shows the features the classifiers had access to. These include character-level features (number of punctuation marks; proportion of punctuation marks; proportion of characters; proportion of digits), word-level features (unigrams, bigrams, and trigrams appearing more than once in the database; number of word types; number of word tokens; lexical diversity; average word length; average word log frequency; proportion of first person pronouns), and sentence-level features (number of sentences per message; average sentence length). This led to approximately 11,600 features, most of which were the unigrams, bigrams, and trigrams (only 42 were features other than these, as shown in Table 3). Some of these shallow features are coarse measures of more complex properties. For example, first person pronouns index self-reference, which is thought to decrease during deception as the deceiver puts more psychological distance between herself and the message (e.g., see [15]).

Feature type	Description	#	Implementation	Sample calculation
punctuation marks	? ! . ; : ,	6	frequency of mark	'c'mon!' = 1 !
characters	Letters a, b, c…z, all digits, all punctuation marks	28	#/ # character tokens	(# digits)/(total # letters, digits, punctuation marks)
n-grams	unigrams, bigrams, & trigrams appearing more than once in the database	varies	frequency of n-gram	'BEGIN+please' appears once in 'please read this'
word types	number of word types	1	# word types	'the penguin ate the fish' = 4
word tokens	number of word tokens	1	# word tokens	'the penguin ate the fish' = 5
lexical diversity	word type to word token ratio	1	# word types / # word tokens	'the penguin ate the fish' ≈ 4/5
average word length	average number of characters per word	1	# characters/ # word tokens	'the penguin ate the fish' = 4
average word log frequency	average of the log of the normalized frequency for each word in the message that appears more than once in the database	1	$\frac{\sum_{w \in msg} \log(\frac{freq(w)}{\sum_{d \in database} freq(d)})}{\#\ \text{word tokens in msg}}$	same as implementation
1st person pronouns	*I, me, my, mine, we, us, our, ours, myself, ourselves*	1	# 1st person pro/ # word tokens	'we saw penguins' ≈ 1/3
sentences	number of sentences	1	# sentences	"what did you see? We saw penguins" = 2
average sentence length	average number of words per sentence	1	# word tokens in msg/ # sentences in msg	"what did you see? We saw penguins" ≈ 7/2

TABLE 3: Linguistic features used by classifiers. Note that for all proportion calculations, a smoothing constant (0.5) was added to the raw counts. Note also that lexical diversity values range between 0 and 1, with higher values indicating more diverse usage (each word appears around once). In addition, all bigrams and trigrams include begin-message (BEGIN) and end-message (END) markers.

3.2 Classifier performance

For each classifier we examined, we used 10-fold cross validation, such that the classifier was trained on the interpretations for 90% of the messages and tested on its predictions of the remaining 10% of the messages, with this process repeated 10 times (for each of the 10 folds). The results reported in table 4 represent the ability of each classifier to predict the correct interpretation for a message, averaged over all messages from the eight tone types. The first classifier selected was the Naïve Bayes classifier, which uses all available features when making its decision. This contrasts with the second classifier selected, the Sparse Multinomial Logistic Regression (SMLR) classifier [37], which uses regression analysis to identify classifier features that are particularly useful for detecting each tone type. In particular, not all features may be useful for each tone type, and this analysis allows us to downweight and possibly remove the features that are less discriminative. Two parameters in the SMLR classifier are λ, which determines how strongly the classifier prefers to rely on a small number of features (i.e., by giving those features non-zero weight), and r, the number of regression rounds. The results in table 4 are from the SMLR classifier that performed the best, using $\lambda = 0.05$ and r = 3.

	Complete	Filtered
# Messages	4,157	1,862
Baseline: 1 in 8	0.125	0.125
Baseline: most frequent in training set	0.137	0.172
Naïve Bayes	0.585	0.655
SMLR (λ = 0.05, r = 3)	0.586	0.704

TABLE 4: Summary of classifier performance across the complete and filtered datasets, including the number of messages in the dataset, two baseline performance measures, and classifier performance.

From these results, we can make several striking observations. First, both classifiers are doing quite well compared to the best baseline, no matter which dataset they use – they are 4.3 times better than the best baseline on the complete dataset (0.585 or 0.586 vs. 0.137) and 3.8 – 4.1 times better than the best baseline on the filtered dataset (0.655 or 0.704 vs. 0.172). This suggests that this kind of dataset is very useful for developing tone detection systems – even fairly small datasets can yield good performance. This relates to the second observation: The SMLR classifier trained on the filtered dataset is very close to human performance already (human: 0.744, SMLR: 0.704), even when using shallow linguistic features. If we want to understand how humans interpret message tone so that we can develop software that will automatically identify unintended tones, these data suggest that even very shallow features may be useful. Third, we can see the importance of using quality messages, even at the expense of the quantity of messages. In particular, while there is equivalent performance by both classifiers on the complete dataset, both improve when using the filtered dataset, with the SMLR improving the most (from 0.586 to 0.704). This is true despite the filtered dataset having approximately a third the number of messages that the complete dataset has.

We also note that the performance of both classifiers is similar to human performance, in that some tone types are more difficult than others. Figures 4 and 5 show confusion matrices for each classifier trained on the filtered dataset.

	deception	politeness	rudeness	embarrassment	confidence	disbelief	formality	persuading	total messages
deception	**0.32**	0.02	0.17	0.08	0.12	0.17	0.00	0.13	163
politeness	0.02	**0.58**	0.21	0.03	0.01	0.05	0.00	0.09	214
rudeness	0.01	0.03	**0.78**	0.03	0.02	0.08	0.00	0.05	321
embarrassment	0.02	0.00	0.10	**0.72**	0.06	0.09	0.00	0.02	246
confidence	0.01	0.02	0.06	0.04	**0.74**	0.09	0.00	0.05	264
disbelief	0.01	0.01	0.09	0.04	0.05	**0.77**	0.00	0.02	300
formality	0.02	0.32	0.28	0.03	0.02	0.06	**0.06**	0.21	98
persuading	0.02	0.04	0.15	0.01	0.05	0.03	0.00	**0.72**	256
									1862

FIGURE 4: Naïve Bayes confusion matrix for the eight tone types investigated. The rows represent the intended tone, while the columns represent the interpreted tones. The bolded diagonal indicates the percentage of correct predictions for each tone type. The total number of messages for each tone type is shown in the rightmost column.

	deception	politeness	rudeness	embarrassment	confidence	disbelief	formality	persuading	total messages
deception	**0.45**	0.00	0.14	0.14	0.10	0.12	0.00	0.05	163
politeness	0.02	**0.58**	0.19	0.05	0.00	0.08	0.02	0.08	214
rudeness	0.02	0.01	**0.77**	0.01	0.01	0.11	0.00	0.07	321
embarrassment	0.03	0.03	0.01	**0.81**	0.05	0.05	0.00	0.01	246
confidence	0.05	0.04	0.00	0.10	**0.79**	0.02	0.00	0.01	264
disbelief	0.06	0.00	0.06	0.04	0.04	**0.78**	0.00	0.02	300
formality	0.04	0.16	0.12	0.04	0.04	0.08	**0.40**	0.12	98
persuading	0.01	0.05	0.06	0.00	0.11	0.03	0.05	**0.69**	256
									1862

FIGURE 5: SMLR confusion matrix for the eight tone types investigated. The rows represent the intended tone, while the columns represent the interpreted tones. The bolded diagonal indicates the percentage of correct predictions for each tone type. The total number of messages for each tone type is shown in the rightmost column.

Similar to the human confusion matrix of figure 3, these figures show the likelihood that a message will be interpreted as a specific tone (in the columns), given that it has been generated with that specific tone in mind (in the rows), averaged over messages. In other words, these figures show the conditional probability distribution *p(predicted l generated*) for the Naïve Bayes classifier (figure 4) and the SMLR classifier (figure 5). The diagonal probabilities indicate how often a message's tone was correctly predicted for each tone type. The total number of messages for each tone type is shown in the rightmost column.

We can observe some similarities in the performance of both classifiers. Similar to humans, both classifiers struggle with deception and formality. Also similar to humans, the precision of formality is very good (Naïve Bayes: .06/Σ(formality column) = 1.00, SMLR: .40//Σ(formality column) = 0.85), as is the precision of deception (Naïve Bayes: .32/Σ(deception column) = 0.74, SMLR: .45//Σ(deception column) = 0.66). However, we also see some non-human confusions in both classifiers. Unlike humans, who mostly confuse formality with politeness and persuading, we find that both classifiers are much more variable in their formality confusions – both often confuse formality with rudeness (Naïve Bayes: 0.28, SMLR: 0.12), for example. A similar non-human behavior occurs with politeness, which is often confused with rudeness (Naïve Bayes: 0.21, SMLR: 0.19), and rarely with formality (Naïve Bayes: 0.00, SMLR: 0.02). Deception is also more often confused with rudeness (Naïve Bayes: 0.17, SMLR: 0.14), unlike what humans do (Humans, figure 3: 0.05).

We can think of two potential reasons for this non-human behavior. First, it may be that there is some kind of default in both classifiers to the most frequent message in the training set when there is uncertainty (rudeness is the most frequent tone in the filtered dataset, at 17.2%). In fact, if we look at the confusions for both classifiers with rudeness in the complete dataset (where rudeness is still the most frequent tone, but now is only 13.7% of the dataset), it turns out there's considerably less confusion with rudeness (Naïve Bayes: deception with rudeness: .09, politeness with rudeness: .09, formality with rudeness: .05; SMLR: deception with rudeness: .09, politeness with rudeness: .09, formality with rudeness: .05). So, this could be a result of the small size of the filtered dataset – in particular, because there are more reliable rude messages in the filtered dataset, it becomes the default prediction, and this is not human-like.

Still, humans have even less confusion with rudeness for these tone types (Human: deception with rudeness: .05, politeness with rudeness: .02, formality with rudeness: .02), which suggests the bias in the training set is only part of the issue. A second cause of the non-human performance could be the particular linguistic features the classifiers are basing their decision on. While shallow linguistic features work fairly well, it may be that we can improve performance to human levels and beyond by tapping into more sophisticated linguistic features.

3.3 Linguistic features for tone

The SMLR classifier offers some insight into what features would be appropriate, as it learns to base its decision on a small number of features. Using $\lambda = 0.05$ and $r= 3$ on the filtered training set, between 497 and 1112 features of the approximately 11,600 available per tone type are given non-zero weight by the classifier. From these, we can see which are strongly weighted, and see if these can suggest useful linguistic features. Table 5 shows a selection of strongly weighted features for each tone type.

Tone	Sample features with strong non-zero weight
deception	i+would+never (3.2), you+have+such (3.1), nope (2.7), umm (2.5), i'm+not (2.0), promise (1.9), would+never (1.9), i'm+just (1.8), i+uh (1.5), i+promise (1.3)
politeness	lovely (4.3), fantastic (3.7), thanks (3.5), love+to (3.0), could+you+please (2.9), thank (2.9), if+you+would (2.4), prettiest (2.3), help+you (2.2), *may-i-take (-1.1)*
rudeness	screw (4.8), filthy (3.5), loser (3.5), annoying (3.5), stupid (3.2), ugly (2.8), fat (2.6), don't+like (2.5), *nice (-1.1), pretty (-1.3), beautiful (-1.7)*
embarrassment	ashamed (3.6), accidentally (3.4), forgot (3.4), awkward (2.7), whoops (2.5), should+have (2.1), oh+no (1.8), 1st person pronouns (1.6), didn't+mean (1.6)
confidence	certain (3.4), really+good (3.3), easy (3.0), i+could (2.8), definitely (2.8), positive (2.7), i'm+sure (2.5), i+look (2.3), i+can+tell (2.3), BEGIN+i+knew (2.2)
disbelief	can't+be (3.8), surprised (3.2), impossible (3.2), didn't+know (3.1), shock (3.0), unreal (2.8), no+way (2.6), outrageous (2.1), # question marks (1.4)
formality	honor (5.5), welcome (3.8), BEGIN+sir (3.5), mrs (3.2), may-i-take (2.5), majesty (2.7), highness (2.2), mr (2.2), may (2.1), madam (1.6), allow (1.6), pardon+me (1.4)
persuading	guarantee (4.8), lets (3.9), just+one (3.0), believe+me (2.8), you+have+to (2.8), fun (2.4), would+look (2.4), trust+me (2.4), think+you+should (2.2), try (1.9)

TABLE 5: A selection of features strongly weighted by the SMLR classifier (λ=0.05, r = 3) for each tone type. The weight given by the classifier is in parentheses after each feature, with negatively weighted features in *italics*.

We discuss some highlights of each tone's features in turn. For deception, instead of finding that first person pronouns are not used, we find that they are used in conjunction with negation words like 'not' and 'never'. We also find indicators of uncertainty ('umm', and 'uh'), and verbs indicating intention ('promise'). For politeness, we find that positive valence words and thanking expressions are positively weighted, in addition to turns of phrase involving some modal verbs ('could', 'would'). Interestingly, we find that the modal 'may' is negatively weighted, presumably because here the classifier is attempting to distinguish between politeness and formality – and 'may' is positively weighted for formality. For rudeness, we find a fairly straightforward pattern of negative valence words being positively weighted and positive valence words being negatively weighted. For embarrassment, we find several words expressing shame or the appearance of an accident ('ashamed', accidentally', awkward', 'whoops', 'oh no'). For confidence, we find many indicators of certainty ('certain', 'definitely', 'positive', 'i'm sure', 'i knew' 'i can tell'), some of which use first person pronouns in them. For disbelief, we find many indicators of surprise ('can't be', 'surprised', 'impossible', 'shock', 'unreal', 'outrageous', the number of question marks), including some that involve negation ('didn't know', 'no way'). For formality, in addition to some fixed formal expressions ('may i', 'pardon me'), we also find several titles ('honor', 'sir', 'mrs', 'majesty', 'highness', 'mr'). For persuading, we see indications of certainty ('guarantee', 'believe me', 'trust me'), positive valence words ('fun'), and coercive expressions ('just', 'you have to', 'think you should', 'try').

Perhaps most notably, the most useful features for these tone types typically are the n-grams, rather than stylometric indicators such as punctuation marks and message length. Given the particular n-grams identified by the SMLR classifier for each tone type, we now have some idea of the more sophisticated syntactic and semantic indicators for each tone. Syntactic classes could continue to include first person pronouns (deception, embarrassment, confidence) while also including negations (deception, embarrassment, disbelief) and modal verbs (politeness, embarrassment, formality, persuading). Semantic classes could include uncertainty (deception), intentions or promises (deception, persuading), social routines (politeness, formality), positive and negative valence (politeness, rudeness, embarrassment, confidence, persuading), shame (embarrassment), surprise (embarrassment, disbelief), accidents (embarrassment, disbelief), certainty (confidence, persuading), titles of address (formality), and coercion (persuading).

The suggested syntactic classes are straightforward enough to extract using defined lists, or perhaps a natural language parser such as the Stanford Parser [38]. The suggested semantic classes pose a more interesting challenge, as they may not be so straightforward to either list or extract. Fortunately, there are some existing tools that may help us make these classes more precise. If we are interested in explicit lists, the Linguistic Inquiry and Word Count database [39] was developed to examine emotional, cognitive, structural, and process components present in language. It includes lists of words that cover positive and negative valence (affective processes: positive and negative emotion), modal verbs (cognitive processes: discrepancy), uncertainty (cognitive processes: tentative), and certainty (cognitive processes: certainty). Similarly, WordNet-Affect [40] was developed to aid in emotion identification research, and includes WordNet classes that correspond to some positive valence words ('joy' class), some negative valence words ('anger', 'disgust', 'fear', and 'sadness' classes), and surprise ('surprise' class).

If we are instead interested in extracting the words and expressions of interest, we may be able to use a machine learning technique called *topic modeling* [41]. Topics are probability distributions over keywords that relate to a cohesive concept such as *food, commerce, casual expressions*, etc. These topics, and the keywords that comprise them, are identified in an unsupervised fashion from a collection of documents. Without any additional information beyond the documents themselves, topic models can use the words contained in the documents to identify both the topics expressed and which topic each word, sentence, or subsection of the document most likely belongs to. For our purposes, this is useful since we are interested in collections of words that correspond to cohesive concepts like *surprise, social routines,* and *coercion*, but which may not have explicit lists of words available. Given a topic model trained over a large enough collection of documents, we may find that a topic model can spontaneously create the list of words associated with some of the concepts of interest. For example, Pearl & Steyvers [36] trained a topic model on the Personal Story Subset of the Spinn3r Blog Dataset [25][26], and this topic model discovered a topic consisting of casual expressions, such as 'oh', 'lol', 'yeah', 'stuff', and 'gonna'. One can easily imagine that such words would not appear in messages with a formal tone. Given a large enough collection of documents, a topic model may thus discover other concepts that are useful for tone detection. Ideally, we would simply train a topic model on messages from the tone database itself, since these are exactly the kind of language text we wish to extract cohesive concepts from. However, we will need to collect significantly more data via Word Sleuth to make this possible, since topic models require large datasets to train on (e.g., the blog entry dataset above had 5.3 million words).

3.4 The ground truth of message tone

As a final note on the success of automatic tone classification, it is worth returning to the issue of the ground truth with respect to message tone. We have assumed so far that the ground truth of a message's tone is the generator's intended tone. While this seems a reasonable approach, we also encountered a situation where the majority of interpreters agree on a message's tone and it is *not* the generator's intended tone. Some examples of this are shown in Table 6, where it seems that majority opinion has converged on a tone other than the intended tone. In this case, is the "true" tone the intended tone or the majority-perceived tone? It may be that it is more reasonable to assume that the generator made a mistake, and instead that the majority-perceived tone is the true message tone. We note that this would differ from our current implementation, where the ground truth for messages in the complete dataset was the intended tone, and only messages where there was majority agreement on the intended tone were included in the filtered dataset. Instead, we could let the true tone be the majority-perceived tone, whatever that tone may be.

Intended tone	Perceived tone	Message
formality	politeness (0.80)	"would you mind please pushing my swing ?"
embarrassment	disbelief (0.70)	"I can't believe John stood me up AGAIN, on our anniversary too."
deception	rudeness (0.67)	"What? I'm not wearing a purple shirt. Your eyes are broken."
confidence	persuading (0.63)	"You should go out there and be yourself in front of others!"

TABLE 6: Sample messages from the current tone database, where the majority of interpreters perceived one tone, even though the generator intended a different tone. The percentage of interpreters identifying the perceived tone are shown in parentheses after the perceived tone.

A related issue is that we have forced participants to choose a single tone to express and perceive, when in fact messages may be more naturally viewed as a mixture of tones, some more strongly expressed than others. This would account for the examples in Table 6, for instance – in each case, the perceived tone and the intended tone are both likely expressed in the message. It's simply that the perceivers disagree with the generator about which tone is expressed more strongly. A future implementation of the Word Sleuth game could allow perceivers to indicate if a message expresses multiple tones, as well as indicating which tones are more strongly expressed. This more nuanced information could then be used to train tone identification systems.

4. CONCLUSION

We have examined the problem of tone identification, viewing it as the expression of a mental state in language text. Aware that there are few existing reliable resources of language text annotated with tone data, we described a methodology for creating such a database using a game with a purpose. We subsequently demonstrated the utility of this database, even though it is currently a small-scale one, on the problem of tone identification. Using two machine learning classifiers that operate over shallow linguistic features, we were able to obtain near human-level identification performance once we applied a simple wisdom of the crowds filter on the dataset. We also discussed some future directions for linguistic features as well as other implementations of the database that may yield better natural language processing performance. Given these initial positive results, the future of the "tone checker" seems promising.

5. ACKNOWLEDGEMENTS

We would like to thank Lawrence Phillips, Galina Tucker, Shannon Stanton, Joseph Nunn, Uma Patel, members of the Computation of Language Laboratory at UC Irvine, and audiences at the Emotions Workshop at NAACL 2010 and the Artificial Intelligence and Machine Learning Series at UC Irvine in 2012.

6. REFERENCES

[1] B. Pang and L. Lee. "Opinion Mining and Sentiment Analysis". *Foundations and Trends in Information Retrieval,* vol. 2(1-2), pp. 1-135, 2008.

[2] A. Abbasi. “Affect intensity analysis of dark web forums,” in Proceedings of Intelligence and Security Informatics (ISI), 2007, pp. 282-288.

[3] A. Agarwal, F. Biadsy, and K. Mckeown. “Contextual Phrase-Level Polarity Analysis using Lexical Affect Scoring and Syntactic N-grams,” in Proceedings of the 12th Conference of the European Chapter of the ACL, 2009, pp. 24-32.

[4] K. Dave, S. Lawrence, and D. Pennock. “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” in Proceedings of WWW, 2003, pp. 519-528.

[5] S. Greene and P. Resnik. “More Than Words: Syntactic Packaging and Implicit Sentiment,” in Proceedings of NAACL, 2009.

[6] A. Kennedy and D. Inkpen, D. “Sentiment classification of movie reviews using contextual valence shifters”. *Computational Intelligence,* vol. 22, pp. 110-125, 2006.

[7] B. Pang, L. Lee, and S. Vaithyanathan. “Thumbs up? Sentiment Classification using Machine Learning Techniques,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79-86.

[8] P. Turney. 2002. “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” in Proceedings of the Association for Computational Linguistics (ACL), 2002, pp. 417-424.

[9] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. “Learning subjective language”. *Computational Linguistics*, vol. 30, pp. 277-308, 2004.

[10] C. Alm, D. Roth, and R. Sproat. “Emotions from text: Machine learning for text-based emotion prediction,” in Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 2005.

[11] P. Anand, J. King, J. Boyd-Graber, E. Wagner, C. Martell, D. Oard, and P. Resnik, "Believe Me -- We Can Do This! Annotating Persuasive Acts in Blog Text", in Proceedings of the AAAI Workshop on Computational Models of Natural Argument, 2011.

[12] P. Subasic and A. Huettner. “Affect analysis of text using fuzzy semantic typing”. *IEEE Transactions on Fuzzy Systems*, vol. 9, pp. 483-496, 2001.

[13] E. Hardisty, J. Boyd-Graber, and P. Resnik. “Modeling Perspective using Adaptor Grammars,” in Proceedings of Empirical Methods in Natural Language Processing, 2010.

[14] W. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. “Which side are you on? Identifying perspectives at the document and sentence levels,” in Proceedings of the Conference on Natural Language Learning (CoNLL), 2006. Internet: https://sites.google.com/site/weihaolinatcmu/data

[15] L. Anolli, M. Balconi, and R. Ciceri. “Deceptive Miscommunication Theory (DeMiT): A New Model for the Analysis of Deceptive Communication,” in Say not to say: new perspectives on miscommunication. L. Anolli, R. Ciceri, and G. Rivs, Ed. IOS Press, 2002, pp. 73-100.

[16] S. Gupta and D. Skillicorn. 2006. “Improving a Textual Deception Detection Model,” in Proceedings of the conference of the Center for Advanced Studies on Collaborative research, 2006.

[17] R. Mihalcea and C. Strapparava. “The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language,” in Proceedings of the Association for Computational Linguistics (ACL), 2009, pp. 309-312.

[18] L. Zhou, J. Burgoon, J. Nunamaker, and D. Twitchell. “Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication”. *Group Decision and Negotiation*, vol. 13, pp. 81-106, 2004.

[19] L. Zhou and Y. Sung. 2008. “Cues to deception in online Chinese groups,” in Proceedings of the 41st Annual Hawaii international Conference on System Sciences, 2008, pp. 146-151.

[20] A. Kosorukoff. “Human-based Genetic Algorithm, “ in IEEE Transactions on Systems, Man, and Cybernetics (SMC), 2001, pp. 3464-3469.

[21] L. von Ahn. “Games With A Purpose”. IEEE Computer Magazine (June, 2006), pp. 96-98.

[22] L. von Ahn and L. Dabbish. “Labeling Images with a Computer Game,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Association for Computing Machinery), 2004, pp. 319-326.

[23] L. von Ahn, M. Kedia, and M. Blum. 2006. “Verbosity: A Game for Collecting Common-Sense Facts,” in proceedings of the SIGCHI conference on Human Factors in computing systems, 2006.

[24] D. Graff & C. Cieri. “English Gigaword.” Linguistic Data Consortium, Internet: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T05, 2003. [Mar 16, 2012].

[25] A. Gordon, A. “Story management technologies for organizational learning,” in Proceedings of the International Conference on Knowledge Management Graz, 2008. Internet: http://ict.usc.edu/files/publications/ IKNOW08.PDF [Feb 10, 2012].

[26] K. Burton, A. Java, and I. Soboroff. “The ICWSM 2009 Spinn3r Dataset,” in Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM), 2009. Internet: http://www.icwsm.org/ data/ [Feb 10 2012].

[27] M. Diab, B. Dorr, L. Levin, T. Mitamura, R. Passonneau, O. Rambow, and L. Ramshaw. “Language Understanding Annotation Corpus”, Linguistic Data Consortium, 2009. Internet: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2009T10 [Mar 16 2012].

[28] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. “Effects of Age and Gender on Blogging”, in Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, 2006.

[29] M. Steyvers, M. Lee, B. Miller, and P. Hemmer. “The Wisdom of Crowds in the Recollection of Order Information,” In Advances in Neural Information Processing Systems, 2009.

[30] B. Turner and M. Steyvers. “A Wisdom of the Crowd Approach to Forecasting,” in Proceedings of the 2nd NIPS workshop on Computational Social Science and the Wisdom of Crowds, 2011.

[31] S. Yi, M. Steyvers, and M. Lee. “The Wisdom of Crowds in Combinatorial Problems.” *Cognitive Science*, to appear 2012.

[32] M. Lee, M. Steyvers, M. de Young, and B. Miller. “Inferring expertise in knowledge and prediction ranking tasks”. *Topics in Cognitive Science*, to appear 2012.

[33] R. Snow, B. O’Connor, D. Jurafsky, and A. Ng. “Cheap and Fast - But is it Good? Evaluating Non- Expert Annotations for Natural Language Tasks,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2008, pp. 254-263.

[34] S. Hacker and L. von Ahn. “Matchin: Eliciting User Preferences with an Online Game,” in Proceedings of ACM Conference on Human Factors in Computing Systems, 2009, pp 1207-1216.

[35] E. Law and L. von Ahn. “Input-Agreement: A New Mechanism for Collecting Data Using Human Computation Games,” in Proceedings of ACM Conference on Human Factors in Computing Systems, 2009, pp 1197-1206.

[36] L. Pearl and M. Steyvers. “Detecting authorship deception: A supervised machine learning approach using author writeprints”. *Literary and Linguistic Computing*., 2012. doi: 10.1093/llc/fqs003.

[37] B. Krishnapuram, M. Figueiredo, L. Carin, and A. Hartemink. “Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 957-968, 2005.

[38] D. Klein and C. Manning. “Accurate Unlexicalized Parsing,” in Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL), 2003, pp. 423-430.

[39] J. Pennebaker and M. Francis. Linguistic Inquiry and Word Count, 1st edition. Lawrence Erlbaum, 1999.

[40] C. Strapparava and A. Valitutti, "WordNet-Affect: an affective extension of WordNet," in the Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), 2004, pp. 1083-1086.

[41] T. Griffiths, and M. Steyvers. “Finding scientific topics”. *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

INSTRUCTIONS TO CONTRIBUTORS

Computational linguistics is an interdisciplinary field dealing with the statistical and/or rule-based modeling of natural language from a computational perspective. Today, computational language acquisition stands as one of the most fundamental, beguiling, and surprisingly open questions for computer science. With the aims to provide a scientific forum where computer scientists, experts in artificial intelligence, mathematicians, logicians, cognitive scientists, cognitive psychologists, psycholinguists, anthropologists and neuroscientists can present research studies, International Journal of Computational Linguistics (IJCL) publish papers that describe state of the art techniques, scientific research studies and results in computational linguistics in general but on theoretical linguistics, psycholinguistics, natural language processing, grammatical inference, machine learning and cognitive science computational models of linguistic theorizing: standard and enriched context free models, principles and parameters models, optimality theory and researchers working within the minimalist program, and other approaches. IJCL is a peer review journal and a bi-monthly journal.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCL.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Started with Volume 4, 2013, IJCL aim to appear with more focused issues related to computational linguistics studies. Besides normal publications, IJCL intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

IJCL List of Topics:
The realm of International Journal of Computational Linguistics (IJCL) extends, but not limited, to the following:

- Computational Linguistics
- Computational Models
- Computational Theories
- Corpus Linguistics
- Formal Linguistics-Theoretic and Grammar Induction
- Information Retrieval and Extraction
- Language Generation
- Language Learning
- Linguistics Modeling Techniques
- Linguistics Theories
- Machine Translation
- Models of Language Change and its Effect on Lingui
- Models that Address the Acquisition of Word-order
- Models that Combine Linguistics Parsing
- Models that Employ Statistical/probabilistic Gramm
- Models that Employ Techniques from machine learning
- Natural Language Processing
- Quantitative Linguistics
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Web Information

CALL FOR PAPERS

Volume: 4 - **Issue:** 2

i. Paper Submission: October 31, 2013 **ii. Author Notification:** November 30, 2013

iii. Issue Publication: December 2013

CONTACT INFORMATION

Computer Science Journals Sdn BhD

B-5-8 Plaza Mont Kiara, Mont Kiara

50480, Kuala Lumpur, MALAYSIA

Phone: 006 03 6204 5627

Fax: 006 03 6204 5628

Email: cscpress@cscjournals.org

COMPUTER SCIENCE JOURNALS SDN BHD
B-5-8 PLAZA MONT KIARA
MONT KIARA
50480, KUALA LUMPUR
MALAYSIA

PHONE: 006 03 6207 1607
006 03 2782 6991

FAX: 006 03 6207 1697
EMAIL: cscpress@cscjournals.org