That's What She Said: Double Entendre Identification

Chloé Kiddon and **Yuriy Brun**
Computer Science & Engineering
University of Washington
Seattle WA 98195-2350
{chloe,brun}@cs.washington.edu

Abstract

Humor identification is a hard natural language understanding problem. We identify a subproblem — the "that's what she said" problem — with two distinguishing characteristics: (1) use of nouns that are euphemisms for sexually explicit nouns and (2) structure common in the erotic domain. We address this problem in a classification approach that includes features that model those two characteristics. Experiments on web data demonstrate that our approach improves precision by 12% over baseline techniques that use only word-based features.

1 Introduction

"That's what she said" is a well-known family of jokes, recently repopularized by the television show "The Office" (Daniels et al., 2005). The jokes consist of saying "that's what she said" after someone else utters a statement in a non-sexual context that could also have been used in a sexual context. For example, if Aaron refers to his late-evening basketball practice, saying "I was trying all night, but I just could not get it in!", Betty could utter "that's what she said", completing the joke. While somewhat juvenile, this joke presents an interesting natural language understanding problem.

A "that's what she said" (TWSS) joke is a type of double entendre. A *double entendre*, or *adianoeta*, is an expression that can be understood in two different ways: an innocuous, straightforward way, given the context, and a risqué way that indirectly alludes to a different, indecent context. To our knowledge, related research has not studied the task of identifying double entendres in text or speech. The task is complex and would require both deep semantic and cultural understanding to recognize the vast array of double entendres. We focus on a subtask of double entendre identification: TWSS recognition. We say a sentence is a TWSS if it is funny to follow that sentence with "that's what she said".

We frame the problem of TWSS recognition as a type of metaphor identification. A metaphor is a figure of speech that creates an analogical mapping between two conceptual domains so that the terminology of one (*source*) domain can be used to describe situations and objects in the other (*target*) domain. Usage of the source domain's terminology in the source domain is *literal* and is *nonliteral* in the target domain. Metaphor identification systems seek to differentiate between literal and nonliteral expressions. Some computational approaches to metaphor identification learn selectional preferences of words in multiple domains to help identify nonliteral usage (Mason, 2004; Shutova, 2010). Other approaches train support vector machine (SVM) models on labeled training data to distinguish metaphoric language from literal language (Pasanek and Sculley, 2008).

TWSSs also represent mappings between two domains: the innocuous source domain and an erotic target domain. Therefore, we can apply methods from metaphor identification to TWSS identification. In particular, we (1) compare the adjectival selectional preferences of sexually explicit nouns to those of other nouns to determine which nouns may be euphemisms for sexually explicit nouns and (2)

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 89–94,
Portland, Oregon, June 19-24, 2011.

examine the relationship between structures in the erotic domain and nonerotic contexts. We present a novel approach — Double Entendre via Noun Transfer (DEviaNT) — that applies metaphor identification techniques to solving the double entendre problem and evaluate it on the TWSS problem. DEviaNT classifies individual sentences as either funny if followed by "that's what she said" or not, which is a type of automatic humor recognition (Mihalcea and Strapparava, 2005; Mihalcea and Pulman, 2007).

We argue that in the TWSS domain, high precision is important, while low recall may be tolerated. In experiments on nearly 21K sentences, we find that DEviaNT has 12% higher precision than that of baseline classifiers that use n-gram TWSS models.

The rest of this paper is structured as follows: Section 2 will outline the characteristics of the TWSS problem that we leverage in our approach. Section 3 will describe the DEviaNT approach. Section 4 will evaluate DEviaNT on the TWSS problem. Finally, Section 5 will summarize our contributions.

2 The TWSS Problem

We observe two facts about the TWSS problem. First, sentences with nouns that are euphemisms for sexually explicit nouns are more likely to be TWSSs. For example, containing the noun "banana" makes a sentence more likely to be a TWSS than containing the noun "door". Second, TWSSs share common structure with sentences in the erotic domain. For example, a sentence of the form "[subject] stuck [object] in" or "[subject] could eat [object] all day" is more likely to be a TWSS than not. Thus, we hypothesize that machine learning with euphemism- and structure-based features is a promising approach to solving the TWSS problem. Accordingly, apart from a few basic features that define a TWSS joke (e.g., short sentence), all of our approach's lexical features model a metaphorical mapping to objects and structures in the erotic domain.

Part of TWSS identification is recognizing that the source context in which the potential TWSS is uttered is not in an erotic one. If it is, then the mapping to the erotic domain is the identity and the statement is not a TWSS. In this paper, we assume all test instances are from nonerotic domains and leave the classification of erotic and nonerotic contexts to future work.

There are two interesting and important aspects of the TWSS problem that make solving it difficult. First, many domains in which a TWSS classifier could be applied value high precision significantly more than high recall. For example, in a social setting, the cost of saying "that's what she said" inappropriately is high, whereas the cost of not saying it when it might have been appropriate is negligible. For another example, in automated public tagging of twitter and facebook data, false positives are considered spam and violate usage policies, whereas false negatives go unnoticed. Second, the overwhelming majority of everyday sentences are not TWSSs, making achieving high precision even more difficult. In this paper, we strive specifically to achieve high precision but are willing to sacrifice recall.

3 The DEviaNT Approach

The TWSS problem has two identifying characteristics: (1) TWSSs are likely to contain nouns that are euphemisms for sexually explicit nouns and (2) TWSSs share common structure with sentences in the erotic domain. Our approach to solving the TWSS problem is centered around an SVM model that uses features designed to model those characteristics. We call our approach Double Entendre via Noun Transfer, or the DEviaNT approach.

We will use features that build on corpus statistics computed for known erotic words, and their lexical contexts, as described in the rest of this section.

3.1 Data and word classes

Let SN be an open set of sexually explicit nouns. We manually approximated SN with a set of 76 nouns that are predominantly used in sexual contexts. We clustered the nouns into 9 categories based on which sexual object, body part, or participant they identify. Let $SN^- \subset SN$ be the set of sexually explicit nouns that are likely targets for euphemism. We did not consider euphemisms for people since they rarely, if ever, are used in TWSS jokes. In our approximation, $\left|SN^-\right| = 61$. Let BP be an open set of body-part nouns. Our approximation contains 98 body parts.

DEviaNT uses two corpora. The erotica corpus consists of 1.5M sentences from the erotica section

of textfiles.com/sex/EROTICA. We removed headers, footers, URLs, and unparseable text. The Brown corpus (Francis and Kucera, 1979) is 57K sentences that represent standard (nonerotic) literature. We tagged the erotica corpus with the Stanford Parser (Toutanova and Manning, 2000; Toutanova et al., 2003); the Brown corpus is already tagged. To make the corpora more generic, we replaced all numbers with the **CD** tag, all proper nouns with the **NNP** tag, all nouns $\in SN$ with an **SN** tag, and all nouns $\notin BP$ with the **NN** tag. We ignored determiners and punctuation.

3.2 Word- and phrase-level analysis

We define three functions to measure how closely related a noun, an adjective, and a verb phrase are to the erotica domain.

1. The **noun sexiness** function $NS(n)$ is a real-valued measure of the maximum similarity a noun $n \notin SN$ has to each of the nouns $\in SN^-$. For each noun, let the *adjective count vector* be the vector of the absolute frequencies of each adjective that modifies the noun in the union of the erotica and the Brown corpora. We define $NS(n)$ to be the maximum cosine similarity, over each noun $\in SN^-$, using term frequency-inverse document frequency (tf-idf) weights of the nouns' adjective count vectors. For nouns that occurred fewer that 200 times, occurred fewer than 50 times with adjectives, or were associated with 3 times as many adjectives that never occurred with nouns in SN than adjectives that did, $NS(n) = 10^{-7}$ (smaller than all recorded similarities). Example nouns with high NS are "rod" and "meat".

2. The **adjective sexiness** function $AS(a)$ is a real-valued measure of how likely an adjective a is to modify a noun $\in SN$. We define $AS(a)$ to be the relative frequency of a in sentences in the erotica corpus that contain at least one noun $\in SN$. Example adjectives with high AS are "hot" and "wet".

3. The **verb sexiness** function $VS(\mathbf{v})$ is a real-valued measure of how much more likely a verb phrase $\mathbf{v}$ is to appear in an erotic context than a nonerotic one. Let S_E be the set of sentences in the erotica corpus that contain nouns $\in SN$. Let S_B be the set of all sentences in the Brown corpus. Given a sentence s containing a verb v, the verb phrase $\mathbf{v}$ is the contiguous substring of the sentence that contains v and is bordered on each side by the closest noun or one of the set of pronouns $\{I, you, it, me\}$. (If neither a noun nor none of the pronouns occur on a side of the verb, v itself is an endpoint of $\mathbf{v}$.)

To define $VS(\mathbf{v})$, we approximate the probabilities of $\mathbf{v}$ appearing in an erotic and a nonerotic context with counts in S_E and S_B, respectively. We normalize the counts in S_B such that $P(s \in S_E) = P(s \in S_B)$. Let $VS(\mathbf{v})$ be the probability that $(\mathbf{v} \in s) \implies (s$ is in an erotic context). Then,

$$\begin{aligned} VS(\mathbf{v}) &= P(s \in S_E | \mathbf{v} \in s) \\ &= \frac{P(\mathbf{v} \in s | s \in S_E) P(s \in S_E)}{P(\mathbf{v} \in s)}. \end{aligned}$$

Intuitively, the verb sexiness is a measure of how likely the action described in a sentence could be an action (via some metaphoric mapping) to an action in an erotic context.

3.3 Features

DEviaNT uses the following features to identify potential mappings of a sentence s into the erotic domain, organized into two categories: NOUN EUPHEMISMS and STRUCTURAL ELEMENTS.

NOUN EUPHEMISMS:

- (boolean) does s contain a noun $\in SN$?,
- (boolean) does s contain a noun $\in BP$?,
- (boolean) does s contain a noun n such that $NS(n) = 10^{-7}$,
- (real) average $NS(n)$, for all nouns $n \in s$ such that $n \notin SN \cup BP$,

STRUCTURAL ELEMENTS:

- (boolean) does s contain a verb that never occurs in S_E?,
- (boolean) does s contain a verb phrase that never occurs in S_E?,
- (real) average $VS(\mathbf{v})$ over all verb phrases $\mathbf{v} \in s$,
- (real) average $AS(a)$ over all adjectives $a \in s$,
- (boolean) does s contain an adjective a such that a never occurs in a sentence $s \in S_E \cup S_B$ with a noun $\in SN$.

DEviaNT also uses the following features to identify the BASIC STRUCTURE of a TWSS:

- (int) number of non-punctuation tokens,
- (int) number of punctuation tokens,

- ({0, 1, 2+}) for each pronoun and each part-of-speech tag, number of times it occurs in *s*,
- ({noun, proper noun, each of a selected group of pronouns that can be used as subjects (e.g., "she", "it"), other pronoun}) the subject of *s*. (We approximate the subject with the first noun or pronoun.)

3.4 Learning algorithm

DEviaNT uses an SVM classifier from the WEKA machine learning package (Hall et al., 2009) with the features from Section 3.3. In our prototype implementation, DEviaNT uses the default parameter settings and has the option to fit logistic regression curves to the outputs to allow for precision-recall analysis. To minimize false positives, while tolerating false negatives, DEviaNT employs the MetaCost metaclassifier (Domingos, 1999), which uses bagging to reclassify the training data to produce a single cost-sensitive classifier. DEviaNT sets the cost of a false positive to be 100 times that of a false negative.

4 Evaluation

The goal of our evaluation is somewhat unusual. DEviaNT explores a particular approach to solving the TWSS problem: recognizing euphemistic and structural relationships between the source domain and an erotic domain. As such, DEviaNT is at a disadvantage to many potential solutions because DEviaNT does not aggressively explore features specific to TWSSs (e.g., DEviaNT does not use a lexical n-gram model of the TWSS training data). Thus, the goal of our evaluation is not to outperform the baselines in all aspects, but rather to show that by using only euphemism-based and structure-based features, DEviaNT can compete with the baselines, particularly where it matters most, delivering high precision and few false positives.

4.1 Datasets

Our goals for DEviaNT's training data were to (1) include a wide range of negative samples to distinguish TWSSs from arbitrary sentences while (2) keeping negative and positive samples similar enough in language to tackle difficult cases. DEviaNT's positive training data are 2001 quoted sentences from `twssstories.com` (TS), a website of user-submitted TWSS jokes. DEviaNT's negative training data are 2001 sentences from three sources (667 each): `textsfromlastnight.com` (TFLN), a set of user-submitted, typically-racy text messages; `fmylife.com/intimacy` (FML), a set of short (1–2 sentence) user-submitted stories about their love lives; and `wikiquote.org` (WQ), a set of quotations from famous American speakers and films. We did not carefully examine these sources for noise, but given that TWSSs are rare, we assumed these data are sufficiently negative. For testing, we used 262 other TS and 20,700 other TFLN, FML, and WQ sentences (all the data from these sources that were available at the time of the experiments). We cleaned the data by splitting it into individual sentences, capitalizing the first letter of each sentence, tagging it with the Stanford Parser (Toutanova and Manning, 2000; Toutanova et al., 2003), and fixing several tagger errors (e.g., changing the tag of "i" from the foreign word tag **FW** to the correct pronoun tag **PRP**).

4.2 Baselines

Our experiments compare DEviaNT to seven other classifiers: (1) a Naïve Bayes classifier on unigram features, (2) an SVM model trained on unigram features, (3) an SVM model trained on unigram and bigram features, (4–6) MetaCost (Domingos, 1999) (see Section 3.4) versions of (1–3), and (7) a version of DEviaNT that uses just the BASIC STRUCTURE features (as a feature ablation study). The SVM models use the same parameters and kernel function as DEviaNT.

The state-of-the-practice approach to TWSS identification is a naïve Bayes model trained on a unigram model of instances of twitter tweets, some tagged with `#twss` (VandenBos, 2011). While this was the only existing classifier we were able to find, this was not a rigorously approached solution to the problem. In particular, its training data were noisy, partially untaggable, and multilingual. Thus, we reimplemented this approach more rigorously as one of our baselines.

For completeness, we tested whether adding unigram features to DEviaNT improved its performance but found that it did not.

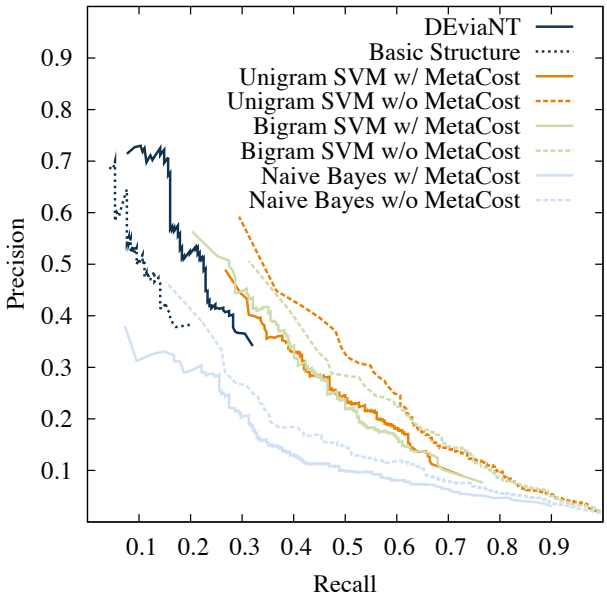

Figure 1: The precision-recall curves for DEviaNT and baseline classifiers on TS, TFLN, FML, and WQ.

4.3 Results

Figure 1 shows the precision-recall curves for DEviaNT and the other seven classifiers. DEviaNT and Basic Structure achieve the highest precisions. The best competitor — Unigram SVM w/o MetaCost — has the maximum precision of 59.2%. In contrast, DEviaNT's precision is over 71.4%. Note that the addition of bigram features yields no improvement in (and can hurt) both precision and recall.

To qualitatively evaluate DEviaNT, we compared those sentences that DEviaNT, Basic Structure, and Unigram SVM w/o MetaCost are most sure are TWSSs. DEviaNT returned 28 such sentences (all tied for most likely to be a TWSS), 20 of which are true positives. However, 2 of the 8 false positives are in fact TWSSs (despite coming from the negative testing data): "Yes give me all the cream and he's gone." and "Yeah but his hole really smells sometimes." Basic Structure was most sure about 16 sentences, 11 of which are true positives. Of these, 7 were also in DEviaNT's most-sure set. However, DEviaNT was also able to identify TWSSs that deal with noun euphemisms (e.g., "Don't you think these buns are a little too big for this meat?"), whereas Basic Structure could not. In contrast, Unigram SVM w/o MetaCost is most sure about 130 sentences, 77 of which are true positives. Note that while DEviaNT has a much lower recall than Unigram SVM w/o MetaCost, it accomplishes our goal of delivering high-precision, while tolerating low recall.

Note that the DEviaNT's precision appears low in large because the testing data is predominantly negative. If DEviaNT classified a randomly selected, balanced subset of the test data, DEviaNT's precision would be 0.995.

5 Contributions

We formally defined the TWSS problem, a subproblem of the double entendre problem. We then identified two characteristics of the TWSS problem — (1) TWSSs are likely to contain nouns that are euphemisms for sexually explicit nouns and (2) TWSSs share common structure with sentences in the erotic domain — that we used to construct DEviaNT, an approach for TWSS classification. DEviaNT identifies euphemism and erotic-domain structure without relying heavily on structural features specific to TWSSs. DEviaNT delivers significantly higher precision than classifiers that use n-gram TWSS models. Our experiments indicate that euphemism- and erotic-domain-structure features contribute to improving the precision of TWSS identification.

While significant future work in improving DEviaNT remains, we have identified two characteristics important to the TWSS problem and demonstrated that an approach based on these characteristics has promise. The technique of metaphorical mapping may be generalized to identify other types of double entendres and other forms of humor.

Acknowledgments

The authors wish to thank Tony Fader and Mark Yatskar for their insights and help with data, Brandon Lucia for his part in coming up with the name DEviaNT, and Luke Zettlemoyer for helpful comments. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant #DGE-0718124 and under Grant #0937060 to the Computing Research Association for the CIFellows Project.

References

Greg Daniels, Ricky Gervais, and Stephen Merchant. 2005. The Office. Television series, the National Broadcasting Company (NBC).

Pedro Domingos. 1999. MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155–164. San Diego, CA, USA.

W. Nelson Francis and Henry Kucera. 1979. *A Standard Corpus of Present-Day Edited American English*. Department of Linguistics, Brown University.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

Zachary J. Mason. 2004. CorMet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.

Rada Mihalcea and Stephen Pulman. 2007. Characterizing humour: An exploration of features in humorous texts. In *Proceedings of the 8th Conference on Intelligent Text Processing and Computational Linguistics (CICLing07)*. Mexico City, Mexico.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP05)*. Vancouver, BC, Canada.

Bradley M. Pasanek and D. Sculley. 2008. Mining millions of metaphors. *Literary and Linguistic Computing*, 23(3).

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT10)*, pages 1029–1037. Los Angeles, CA, USA.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT03)*, pages 252–259. Edmonton, AB, Canada.

Kristina Toutanova and Christopher Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC00)*, pages 63–71. Hong Kong, China.

Ben VandenBos. 2011. Pre-trained "that's what she said" bayes classifier. `http://rubygems.org/gems/twss`.