AMBIGUOUS LEXICAL RESOURCES FOR COMPUTATIONAL HUMOR GENERATION

Alessandro Valitutti
Department of Computer Science and HIIT, University of Helsinki, Helsinki, Finland

Keywords: Ambiguity, Computational Humor, Creative Lexical Resources.

Abstract: The ongoing work presented here is aimed to investigate to what extent it is possible to perform a feasible use of ambiguous texts in computational humor generation. The first core of a lexical database was developed in order to collect ambiguous terms in the English lexicon. Then an exploratory use of the resource for computational humor generation was performed. Finally, three existing prototypes of humor generator were simulated in order to generate different form of humorous messages from the same lexical resource.

1 INTRODUCTION

Humor and creativity are strictly connected. As wittily pointed out by Joel Goodman, "there is a connection between HA HA, and AHA" (Goodman, 1995). In the generation of a joke, a typical creative process consists of the invention of new ways to violate recipients expectation and then induce surprise. A more specific form of creativity is in the discovery of connections that allows the humorist to emphasize ridiculous aspects of people.

Nevertheless in most cases part of the information necessary for the creation of humorous surprise effects is already present in the common sense knowledge and the linguistic use. Creativity in this case consists of the appropriate reuse of pre-existing pieces of knowledge coded in the language.

This paper is focused on linguistic ambiguity. The use of ambiguous texts is a common and effective way to achieve the surprise effect. More specifically, the ongoing work presented here is aimed to investigate to what extent it is possible to perform a feasible use of ambiguous texts in computational humor generation. As a first step, the focus is on the lexical level. The first core of a lexical database, characterized as an extension of WORDNET 3.1 (Fellbaum, 1998), was developed in order to collect ambiguous terms in the English lexicon. Items are defined according to three different possible types of lexical ambiguity (homonymy, homophony and idiomatic ambiguity) and called *double-edged words* (DEW). The database was then called DOUBLE-EDGED WORDNET (**DEWN**).

As a second step, an exploratory use of the resource was started. Three existing prototypes of humor generator were simulated in order to generate different form of humorous messages from the same lexical resource. In this way, the aim is to perform some step toward a more general model of humor generation, in which part of the linguistic knowledge can be reused and extended over the time.

2 BACKGROUND

To date there are only a limited number of researches on the computational generation of humorous texts. Ritchie provides a systematic review of the most remarkable verbal humor generators developed in the last 20 years (Ritchie, 2004). The most remarkable of them are LIBJOG, a program for the generation of light bulb jokes (Raskin and Attardo, 1994), JAPE program producing a specific type of punning riddles (Binsted et al., 1997), and HAHAcronym, a generator of humorous acronyms (Stock and Strapparava, 2002)

3 CHARACTERIZATION OF DOUBLE-EDGED WORDS

The design and development of DEWN is based on the idea of *double-edged word* (from now on called **DEW**), an abstract data structure introduced for modeling a specific type of ambiguous lexical unities. A

Valitutti A..
AMBIGUOUS LEXICAL RESOURCES FOR COMPUTATIONAL HUMOR GENERATION.
DOI: 10.5220/0003882305320535
In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence* (ICAART-2012), pages 532-535
ISBN: 978-989-8425-95-9

humorous DEW is defined as a word with two meanings, one of which is, at the same time, the least common and most interesting one. More specifically, a DEW can be characterized by the following attributes:

- WORD is the lexical unit (e.g. a single word or a phrase).
- AMBIGUITY is a list of two or more "meanings" associated to the WORD.
- DEPTH expresses the different typicality of the two meanings. For example, a two fold ambiguity will be associated to a main meaning (called *surface meaning*, with depth 1) and a secondary meaning (called *hidden meaning*, with depth 2).
- SLANT is a set of additional semantic labels associated to the hidden meaning, and characterizing it as potentially humorous. Slant labels can be used to emphasize the humorous role of hidden meaning. For example, slant labels can be selected in order to evoke ridiculous trait of people.

Two main operations are associated to a database of DEWs: 1) extraction of attribute value of a DEW associated to an input word and 2) selection of the subset of DEWs corresponding to an input slant. The proper indexing of a large database of DEWs according to the slant values is crucial for an efficient retrieval of items for creative applications.

4 RESOURCE DESCRIPTION

The development of DEWN was performed according to three different types described below. Each of them corresponds to a different form of lexical ambiguity: *homonymy*, *homophony*, and *idiomatic ambiguity*.

4.1 Homonymic DEWs

Homonymy is defined as the relation between words that share the same spelling and pronunciation but have different meanings. This is the most typically recognized form of lexical ambiguity and the one employed to define word meanings in a monolingual English dictionary. The term is used here as synonym of *polysemy*, even thought the latter one is often used to indicate words that have at least some feature in common (Blank, 1999). In WordNet each word meaning is represented by a set of synonyms (*synset*) and associated to a specific ID in the database. Each word is associated to one of more *senses* (i.e. ranked synsets). The sense ranking is performed according to their occurrence frequency in a reference corpus annotated according to WordNet senses. So it is natural to identify homonymic DEWs as words in WordNet with at least two senses. The sense number expresses the DEPTH attribute. A list of 24167 DEWs was extracted from WordNet 3.1.

4.2 Homophonic DEWs

Homophony is defined here as the relation between words that are phonetically identical (*complete homophones*) or similar (*partial homophones*) but with different spelling.

The algorithm for the measure of the phonetic distance is a specific implementation of the Levenshtein distance (Levenshtein, 1966). It is based on a sequence of elementary operations applied on the phonetic expression of a word in order to obtain another word. Each step (i.e. application of an operation) is associated to the value of a cost function. The sequence of steps, required to transform the first word in the second one, and corresponding to the minimum total value of cost, defines the distance between two words. Three types of elementary operations are considered: *substitution*, *insertion* and *deletion*.

The cost value associated to the substitution operator was assigned according to the phonetic type, tonic accent, and vowel length. The algorithm reduces the phonetic distance between words to the distance between syllables, and the syllabic distance to the distance between single phonemes.

The information on mapping between words and their phonetic transcription was extracted from the CMU pronouncing dictionary (available at `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`). A measure of the above described phonetic distance was calculated for all pairs of words in WordNet, in order to collect sets of homophones. A number of 5400 total homophonic sets and 23050 partial homophonic sets were filtered.

4.3 Idiomatic DEWs

Idiomatic ambiguity is a specific type of ambiguity between literal and figurative language. Idioms are defined here as multiword expressions whose meaning cannot be inferred by the meaning of the component words. The idiomatic meaning of a word is the meaning associated to the idiom in which the word is included.

A manual annotation of WordNet was performed in order to identify lexical idioms (i.e. idioms consisting of a composed word). The collection includes 3541 WordNet synsets. For each of them, one or more component words were selected. For each idiomatically ambiguous word, the surface meaning (or *literally meaning*) was defined as its first sense in Word-

Net, and the hidden (or *idiomatic meaning*) as the first sense in the idiom in which the word is included.

4.4 Slant Indexing

In order to implement the SLANT attribute, (characterizing potentially "interesting/relevant" meanings for creative/humorous applications), a number of semantic constraints were considered. Semantic constraints can be classified in two categories: 1) *absolute* (i.e. applied to a single meaning) and 2) *relational* (i.e. applied to a couple of meanings of the same word).

A next explorative annotation of previous collected ambiguous terms was performed, exploiting three lexical collections: WORDNET 3.1, WORDNET-DOMAINS (Magnini and Cavaglià, 2000) and a list of positive/negative/polarized words. A first semantic labeling of DEWs was performed taking advantage of WordNet-Domains, and extension of WordNet, in which synsets are tagged according to a list of *semantic domains*. Since the last release of WordNet-Domains is interfaced to WordNet 3.0, the mapping to the 3.1 release was applied.

Other constraints were applied (and additional lists of labeled words employed) in order to emphasize the two following types of semantic opposition).

- **Polarized Words.** A list of positive and negative words collected from the Web and the WordNet-Affect lexical database (Strapparava and Valitutti, 2004) were both employed to filter ambiguous words associated to meanings with opposite values of polarity.
- **Metaphors.** A list of metaphors for people was automatically built exploiting the hypernym hierarchy in WordNet. A list of high-level synsets (called here *metaphor categories*) was defined. The list includes categories such as ANIMAL (see the example in the next section, based on the definition of 'pig'), FOOD and TOOL. The criterion for the selection of DEWs is that the default sense is a descendant, in the hypernym hierarchy, of a metaphor category, and the hidden sense is a descendant of the category PERSON.

5 USE OF DEWN IN HUMOR GENERATORS

As first exploratory use DEWN in computational humor, a few examples are analyzed below. They are obtained through the application of procedures simulating a number of well-known computational humor generators.

5.1 Examples of Punning Riddles

How do you define a pig?
It is a stout-bodied short-legged omnivorous policeman.

In order to obtain this joke, the homonymic DEW "pig" was selected. The definition (in the form of answer) is the gloss of the default meaning (i.e. first WordNet sense of the corresponding noun), in which the word "animal" was substituted by the first synonym ("policeman") of the hidden meaning (i.e. third WordNet sense).

The creation of a punning riddle starting from a "lexical core" is inspired to the JAPE system (Binsted and Ritchie, 1994), in which the joke is generally based on a couple of phonetically similar words. An analogue example is:

Who is a working girl?
A young streetwalker who is employed.

In this case, the definition is obtained though replacing "woman" (in the gloss of the default meaning) with "a young streetwalker" (from the hidden meaning).

5.2 Examples of Funny Acronyms

This type of acronym generation is modeled on the HAHAcronym system (Stock and Strapparava, 2002):

CPU = Celibate Professing Untied

The acronym is generated through the replacement of each word in the original expansion (*Central Processing Unit*) according to phonetic similarity ("processing" vs. "professing") and semantic opposition ("computer" vs. "religion").

The following "hand-made" example, instead, cannot be generated with the present resource because it involve a model of the ambiguity propagated at the phrase level:

IBM = Interpreting Bible Machines
(from the original *International Business Machines*)

5.3 Variation of Familiar Expressions

The following example is based on the FEVER program (Valitutti, 2011):

A chapel a day keeps the malefactor away.

This pun is obtained through two word replacements in which both phonetic similarity and domain slanting (RELIGION) constraints were applied.

Instead the following hand-made expression cannot be generated without a model describing the ambiguity at the sentence level:

An onion a day keeps everyone away.

6 CONCLUSIONS AND FUTURE WORK

Through the development and description of DEWN, this work emphasizes the advantage to use a collection of ambiguous lexicon in computational humor generation. The resource is based on the definition of an abstract data structure (DEW) and aims to simplify and standardize a set of lexical operation employed in existing systems for generating creative text. The applicative examples were selected to support the idea of reuse of available creative operations and their integration through the access to a shared lexical resource.

A crucial aspect in the future development of this type of lexical resources is the indexing of items according to specific semantic dimension, especially when the number of items is enough large to delay the search time.

The sharing of linguistic resources specialized for creative applications and the effort to integrate different specialized humor generators in a more general tool is aimed as a form of adjacent possible. According to this term coined by Stuart Kauffman (Kauffman, 2000), the possible creative achievements available at a given time are based on the existing resources and the shared innovation. The proposed approach is aimed to give a contribution to extend the space of creative possibilities.

ACKNOWLEDGEMENTS

This work has been supported by the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland.

REFERENCES

Binsted, K., Pain, H., and Ritchie, G. (1997). Children's evaluation of computer-generated punning riddles. *Pragmatics and Cognition*, 2(5):305–354.

Binsted, K. and Ritchie, G. (1994). An implemented model of punning riddles. In *Proc. of the 12th National Conference on Artificial Intelligence (AAAI-94)*, Seattle.

Blank, A. (1999). Polysemy in the lexicon. In Eckardt, R. and von Heusinger, K., editors, *Meaning Change – Meaning Variation*, volume 1, pages 11–29.

Fellbaum, C. (1998). *WordNet. An Electronic Lexical Database*. The MIT Press.

Goodman, J. (1995). *1,001 Ways to Add Humor to Your Life and Work*. Health Communications.

Kauffman, S. A. (2000). *Investigations*. OUP.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Magnini, B. and Cavaglià, G. (2000). Integrating subject field codes into wordnet. In *Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*, Athens, Greece.

Raskin, V. and Attardo, S. (1994). Non-literalness and non-bona-fide in language: approaches to formal and computational treatments of humor. *Pragmatics and Cognition*, 2(1):31–69.

Ritchie, G. (2004). *The Linguistic Analysis of Jokes*. Routledge, London.

Stock, O. and Strapparava, C. (2002). HAHAcronym: Humorous agents for humorous acronyms. In *(Stock et al., 2002)*.

Stock, O., Strapparava, C., and Nijholt, A., editors (2002). *Proceedings of the The April Fools Day Workshop on Computational Humour (TWLT20)*, Trento.

Strapparava, C. and Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In *Proc. of 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon.

Valitutti, A. (2011). How many jokes are really funny? towards a new approach to the evaluation of computational humour generators. In *Proc. of 8th International Workshop on Natural Language Processing and Cognitive Science*, Copenhagen.