A Multi-Lingual Dictionary of Dirty Words

Jonas Sjöbergh and Kenji Araki

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

Abstract

We present a multi-lingual dictionary of dirty words. We have collected about 3,200 dirty words in several languages and built a database of these. The language with the most words in the database is English, though there are several hundred dirty words in for instance Japanese too. Words are classified into their general meaning, such as what part of the human anatomy they refer to. Words can also be assigned a nuance label to indicate if it is a cute word used when speaking to children, a very rude word, a clinical word etc. The database is available online and will hopefully be enlarged over time. It has already been used in research on for instance automatic joke generation and emotion detection.

1. Introduction

Dictionaries can be tremendously useful in many language processing tasks, and are also useful sources of information for human readers. One category of words that is often not included in very large amounts in most dictionaries are "dirty words". By that we mean words that are generally not used in polite company, referring to for instance sexually related things, bodily functions, or cuss words and insults.

We have collected dirty words in several languages and built a multi-lingual dictionary linking words in different languages with similar meanings. Other types of annotation is also possible, many words are for instance annotated with the nuance they carry, i.e. if a word is really rude or perhaps a euphemism and so on. Information on whether a word is unambiguously a dirty word or can also have non-dirty meanings can also be added.

This dictionary can of course be used in translation applications, to find appropriate translation candidates for words and phrases that are perhaps hard to find in other dictionaries. A dictionary of dirty words is also useful in many other ways, there have for example been quite a few products launched world wide were the producer later found out that in some of the markets the product name was a dirty word or in a similar way gave bad impression. Such things could be mitigated by having access to a large resource of dirty words in different languages.

The dictionary can also be used in monolingual natural language processing applications where information about dirty words is useful. Three examples of areas where our database has already been used are:

Humor recognition (Sjöbergh and Araki, 2007b). In a machine learning approach to classify texts as either jokes or not, some features based on the presence of dirty words in the text were used. A high presence of dirty words was useful as an indication that the text was a joke.

Humor generation (Sjöbergh and Araki, 2007a; Sjöbergh and Araki, 2008). Dirty words and euphemisms are common in jokes, and are thus useful in automatic joke generation. A system generating rather weak puns was perceived as slightly funnier if the punch line of the pun was a dirty word. Other joke generation methods were based on changing parts of idioms to similar sounding euphemisms (dirty words) for sex etc.

Emotion recognition (Ptaszynski et al., 2007). In a similar way to the humor recognition case, sentences with dirty words tended to carry emotive content in an experiment on emotion detection in Japanese.

Other uses of dirty words, though not our database, in natural language processing include detecting if a message is a flame (Spertus, 1997), and other machine learning approaches to humor recognition (Mihalcea and Strapparava, 2005).

2. Collecting the Dirty Words

We have collected dirty words and short phrases from several different sources and in several different languages to add to our database. These were then annotated manually with various types of information. The original intended use was humor generation and humor recognition in English and Japanese, so these two languages received the most focus.

The single largest source of dirty words was a list collected by George Carlin[1], containing about 2,400 dirty word expressions in English. Most of these are euphemisms, tending towards joke like expressions, for example "trouser anaconda".

For Japanese we extracted all words in the EDICT dictionary (Breen, 1995) marked with the "vulgar" flag, and also added various short lists of dirty words found on the Internet. We also had several native speakers of Japanese simply write down a lot of dirty words that they could come up with by looking at the other words in the list.

We have also found useful information in the Alternative Dictionaries[2], the Swearsaurus[3], and Wikicurse[4], which are

[1] http://www.georgecarlin.com/dirty/2443.html
[2] http://www.notam02.no/~hcholm/altlang/
[3] http://www.insultmonger.com/swearing/
[4] http://www.wikicurse.com/

collections of "bad words" in many languages. There are also many bad words in these resources in other languages that we have not added to our database, mainly because of a lack of native speakers to check if the words are really of the kinds we want. These could of course be added later if one so wishes.
After collecting the dirty words, they have been annotated by hand with different types of information. Not all words are annotated with all types of information yet. Annotation regarding the meaning, nuance, and ambiguity of a word or phrase is possible.

3. Structure

In the dictionary the words are annotated with the following information: how to write the word, how to pronounce the word, the meaning of the word, the nuance of the word, whether the word is ambiguous in the sense that it has non-dirty meanings too, what language the word comes from, and the part of speech of the word. The dictionary also contains many multi-word expressions, though they are treated like one unit and we will refer to these too as "words" in this paper except when talking specifically about the number of words in the expressions.
The only information that is mandatory for a word is how to write it. All other fields can be left unspecified, though so far all words are also annotated with the language they come from. Pronunciation is currently only provided for the Japanese words, for which it can be non-trivial to figure out the reading of the ideographic characters used for writing. The same ideographic character sequence can have several different readings, some of which can be dirty words while others are not.
The meanings are specified by links to special "interlingua" like objects. These describe the general meaning of a word using English (though adding explanations in other languages too is of course also possible). Currently only the general meaning is given, such as what part of the human anatomy a word refers too or that it is some form of fornication. More detailed classifications can be done later if it is found to be necessary for a specific application.
These interlingua meaning objects are also grouped into three general groups: sex related, bodily functions, and insults. The meanings of some words do not fit into any of these three categories, in which case what group the meaning belongs to is left unspecified. An example of an interlingua object is "cuss word interjection" for things such as "dammit".
The nuances of words indicate if a word is a clinical word used for instance in doctor patient conversations, if it is a "cute" word used when speaking to children, if it is a euphemism, or if it is an "extra bad" word (very rude), etc. We have found use for this type of information ourselves in other experiments, for instance in humor generation where really bad words tended to offend rather than entertain, and clinical words did not sound very funny either. This type of information could also be useful for instance when selecting from different translation candidates, so as to find a translation with a similar nuance in the target language.
Nuance can of course be hard to determine for some words. Words can be perceived as very rude by some people and as fairly OK by others. The same word can also be very rude in some contexts and not rude at all in other contexts. Currently we have not made any efforts at more detailed descriptions of nuances, but if there is interest in the future it could be added later. Many words are still unproblematic though, and can be fairly easily annotated with a simple description of their nuance.
The ambiguity field indicates if a word has both dirty meanings and non-dirty meanings. It is possible to just note that both are possible (e.g. "pussy") or if a word is always dirty (e.g. "fuck"), and it is also possibly to specify in more detail if the dirty meaning is much more common than any non-dirty meanings (as perhaps "cock"), or if the word is generally not dirty but can be in special contexts (e.g the words "it" or "there" in many languages). Which meaning is more common can of course in many cases be rather hard to judge, in which case just noting that the word is ambiguous is enough.
Part of speech is currently mostly not given, though the field was added since this information was available in some of the sources we used to build the dictionary and was used in some of our text generation experiments using the data. Other grammatical information could also be useful but is currently not given. For instance, it could be useful to know what forms the mutli-word expressions can take, if they can have parts of the expression modified by adjectives etc. without losing their dirty impression, etc.
The language field simply indicates what language a word comes from. If the same string is a dirty word in several languages, a separate entry is made in the database for each language. The same is true if a word can mean several different dirty things in the same language.
The dictionary is stored in an SQL database. The database has a primitive web interface that allows searching the database, downloading the whole dictionary, adding new words and meanings, and annotating existing words with meanings, nuances, etc.

Language	Words
English	2402
Japanese	397
Swedish	158
Bulgarian	147
Polish	125
Total	3229

Table 1: The number of words and expressions in different languages currently in the database.

4. Statistics

Some statistics showing the contents of the dictionary can be found in Tables 1 to 5. As can be seen in Table 1, the bulk of the words are currently English words. That English has the by far largest amount of words is probably caused by English being the most widely used language both on the Internet (a good source of dirty words) and in natural language research (where one could perhaps expect such resources to show up). Somewhere around 150 words seems

Category	Words
Sex	2652
Bodily Functions	261
Insults	211
Unspecified	105

Table 2: The general grouping of the meanings of the words in the database.

Nuance	Words
Euphemism	1462
Fairly bad word	87
Used normally	70
Children's speak	24
Very bad	22
Clinical	13
Unspecified	1551

Table 3: The nuances of the words.

to be the limit where people who collect dirty words mainly for fun get tired and give up. The differences in number of words between the languages in our database most likely do not reflect any actual cultural differences in the amount of dirty words. It is simply an effect of what purposes we have used the data for and thus what languages we put the most work into collecting dirty words for.

In Table 2, it can be seen that rather unsurprisingly the overwhelming majority of the words are sex related words. This varies a bit between languages though, some of the lists of "dirty words" we have collected have contained mostly insults to be hurled at other people to make them angry. Many of these do have sexual connotations though.

Of the words that have so far been annotated with their nuances, euphemisms are by far the most common, see Table 3, though only about half the words have been annotated so far.

The annotation of the ambiguity of the words has only covered about a third of the data so far, see Table 4. Thus far, about half the words are ambiguous, though a large part of the remaining words are rather long euphemistic expressions that are likely not very ambiguous.

In the final table, Table 5, statistics on the lengths of the expressions is presented. About half the dictionary is made

Ambiguity	Words
Ambiguous, can be either dirty or not	686
Always dirty	475
Ambiguous, not-dirty meaning most common	26
Ambiguous, dirty meaning most common	17
Unspecified	2025

Table 4: The ambiguity of the words.

Length in Words	Expressions
1	1722
2	902
3	398
4	132
5	47
6	20
7	5
8	2
10	1
Multi-Word	1507

Table 5: The lengths of the expressions in the database. Average length is 1.8 words.

up of multi-word expressions, though not many are made up of four or more words. The longest expression so far is "choke the sheriff and wait for the posse to come", which is an English expression for (male) masturbation.

This data gives a general idea of the contents of the dictionary, but one should keep in mind that it is a bit complicated to gather this type of information from such different languages. The Swedish part contains many quite long compound words treated as only one word, while a similar word in English would be a multi-word expression. And Japanese has no space between words at all, so only a quick cursory check of roughly how many "words" a phrase contains was done for the Japanese part.

5. Availability

The dictionary is freely available on the web[5], though the web interface is still very primitive. It is possible to download the whole dictionary, and also to add new words, change or add more annotations to the words already in the database etc.

We plan to extend the database ourselves, both by adding more words to the languages already included but and by adding more languages. Any volunteers are of course also welcome to add more data too. We also plan to improve the web interface.

6. Conclusions

We presented a dictionary of dirty words in several languages. The meanings of the words are linked, so it can be used to find for instance translations of dirty words in other languages. The nuances of the words (really rude, clinical, euphemism, etc.) are also annotated, which can help in selecting an appropriate translation. Words that are ambiguous in the sense that they have other non-dirty meanings too, can also be annotated with this information.

The dictionary contains about 3,000 words and expressions, 2,400 in English, 400 in Japanese, and slightly over 100 words each in Bulgarian, Polish, and Swedish. It is freely available on the Internet, and it is also possible for volunteers to contribute new words to the dictionary.

[5] http://dr-hato.se/projects/dirtywords/

So far, the contents of the dictionary have mainly been used in monolingual applications, for instance humor generation and emotion classification.

Acknowledgements

This work was done as part of a project funded by the Japanese Society for the Promotion of Science (JSPS). We would like to thank some of the anonymous reviewers for interesting suggestions for extending our work. We would also like to thank the volunteers who have contributed dirty words to the dictionary, especially Svetoslav Dankov who also helped out with various practical things.

7. References

Jim Breen. 1995. Building an electronic Japanese-English dictionary. In *Japanese Studies Association of Australia Conference*, Brisbane, Australia.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of HLT/EMNLP*, Vancouver, Canada.

Michal Ptaszynski, Pawel Dybala, Wen Han Shi, Rafal Rzepka, and Kenji Araki. 2007. Lexical analysis of emotiveness in utterances for automatic joke generation. ITE Technical Report, Vol. 31, No. 47, pages 39–42, ME2007-204.

Jonas Sjöbergh and Kenji Araki. 2007a. Automatically creating word-play jokes in japanese. In *Proceedings of NL-178*, pages 91–95, Nagoya, Japan.

Jonas Sjöbergh and Kenji Araki. 2007b. Recognizing humor without recognizing meaning. In Francesco Masulli, Sushmita Mitra, and Gabriella Pasi, editors, *Proceedings of WILF 2007*, volume 4578 of *Lecture Notes in Computer Science*, pages 469–476, Camogli, Italy. Springer.

Jonas Sjöbergh and Kenji Araki. 2008. What is poorly said is a little funny. In *Proceedings of LREC-2008*, Marrakech, Morocco.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Innovative Applications of Artificial Intelligence (IAAI)*, pages 1058–1065, Providence, Rhode Island.