Unnatural language detection

Thomas Lavergne (Jeune Chercheur)

France Telecom R&D
2 av Pierre Marzin
22300 Lannion
thomas.lavergne@rd.francetelecom.com

ABSTRACT. *In the context of web search engines, the escalation between ranking techniques and spamdexing techniques has led to the appearance of faked contents in web pages. If random sequences of keywords are easily detectable, web pages produced by dedicated content generators are a lot more difficult to detect.*
Motivated by search engines applications, we will focus on the problem of automatic unnatural language detection. We will study both syntactical and semantical aspects of this problem, and for both of them we will present probabilistic and symbolic approaches.

RÉSUMÉ. *Dans le contexte des moteurs de recherche sur le web, l'escalade entre les techniques de classement et les techniques de spamdexing a conduit à l'apparition de faux contenus dans les pages web. Si les séquences aléatoires de mots-clés sont facilement détectables, les pages web produites par des générateurs automatiques dédiés sont beaucoup plus difficiles à détecter.*
Motivé par cette application, on se concentrera sur le problème plus général de la détection du catactère peu-naturel d'un texte. On étudiera à la fois les aspects syntaxiques et sémantiques du problème, et pour chacun d'eux on présentera des approches probabilistes et symboliques.

KEYWORDS: *natural language processing, text generator, spamdexing*

MOTS-CLÉS : *traitement automatique des langues, générateur de texte, spamdexing*

1. Introduction

We want to be able to distinguish a computer generated text or a random sequence of words from the real language of humans. We first need to define precisely the notions of natural and unnatural languages. Deciding if a text has been written by a human or not is by essence a subjective task. The only way to surround the subjective aspect of the problem is, like for the Turing test, to use humans themselves as oracles. We call *natural*, a text that a human would consider natural. We call *unnatural* a text that a human would consider artificial.

Many different forms of unnatural text can be found : From the simplest and most artificial, like *word salad* (random sequences of words) to the most elaborated like *content generator*. Some examples of unnatural texts are presented in Figure 1. In order to handle the different forms of unnatural text, we need to consider different techniques of detection and ranking.

This detection has many different applications in information retrieval. For example, in the context of web search engines, automatically generated web pages with no interesting content are created to increase *ranking* of a target web site. These pages constitute what we call a *link farm* ([GYO 05]). This is a form of *spamdexing* : contents only destined to web search engines in order to falsify the ranking of a web site. Such contents need to be detected and filtered.

We present here an overview of the different techniques we plan to investigate for the detection of unnatural language.

2. Text features

There are two aspects of the text that we must consider in order to detect unnatural language : *syntactic* aspects and *semantic* aspects. The syntactic aspect is the internal structure of the text. This structure is guided by the rules of the natural language such as its grammar.[1] The semantic aspect of the text is about its meaning and its consistency. It describes what a reader would understand of the text.

For both of these aspects, we will study two different approaches. Firstly, a *statistical* approach, which uses informations extracted from some natural (or unnatural) examples using techniques like machine learning. This approach considers words as text components, and studies their distribution. The second approach is a *symbolical* one. This approach uses existing knowledge and considers inner informations about words such as their part of speech and how they interact with other words.

[1] Here "grammar" does not refer to a formal grammar.

Word salads :
6708 sports bettingonline sports bettingmarch madnessbasketball bettingncaa bettingsports…

Quotes with keywords :
Cars are not merely *divx* continually perfected mechanical *download* contrivances; since the 1920s nearly *movies* all have been mass-produced to meet a market,…

Machine language :

```
for (i = 0; i < max; i++)\linebreak
  if (tab[i] != NULL)\linebreak
  add\_elem(tab[i]);
```

Content generator :
Notwithstanding the instability of the cosmological parameters (Gauss, 1845), it would be interesting to compute by iterations every random structures. In spite of the non-gaussianity of wavelet transforms (Bessel, 1840), a possible method would be to minimize one maximum of the biased estimates.

Figure 1. *Some example of unnatural texts.*

3. Structural aspects

All natural texts have to respect a lot of rules like grammar rules. The structure produced by these rules can be used to detect unnatural texts.

3.1. *Statistical Approach*

In natural languages, we can exhibit some common characteristics : the more frequent words are *function words* such as determiners, prepositions…like *the, and, a, to,* … . A first simple test is to count the number of occurrences of each words and check if the more frequent ones are mostly function words. This criterion can be used to detect very basic forms of word salad, like sequences of keywords without any function words.

This is a particular case of a more general law of the word distribution : the *Zipf law* (see [ZIP 49]) which says that the frequency of a word is inversely proportional to its distribution rank. Even though this is to be taken with care, we can regard this as a roughly accurate characterization of real data. This could be used to detect more advanced word salad, if the generator is tuned to include function words. Variation around this shall be considered, like looking at the distribution of word length, or consider n-grams instead of words alone.

More advanced statistical methods, like machine learning, can be used to study the structure of a text like in [MAN 99]. *Hidden Markov Models* (see [PAU 90]) are probabilistic models of sequences of elements. A common use of them was to modelize things like sequence of words. We can train an HMM on a manually classified corpus, and then, use it to evaluate the probability of a text to be natural.

3.2. *Symbolic approach*

As opposed to strictly statistical methods, symbolic ones look at the text as linguistic objects. A second approach of structural aspects of the text is to use symbolic methods. A first step is to split the text in labelled lexical units. Different granularities can be used. We can look at each word as one unit and tag it with its *word class*, like *noun*, *verb*, *adjective*... But, it is also possible to split more roughly the text by grouping words in functional groups and label each group with its *function*, like *subject*, *verb*... using *syntactic tagger*.

With this first step we obtain the structure of the text. Then, it is possible to check if this structure looks like the structure of a natural language. These informations about natural languages often come from hand-written grammars or rules, or may be learnt from a set of natural language texts in a mixed approach using statistics.

Another possibility to use the obtained structure, if we have enough input data, is to count how many patterns can be found in the text. This can be viewed as a measure of the *syntactical richness* of the language. Texts generated by dedicated generators will be poor in case of a too small number of rules used for generation.

4. Semantical aspects

When we try to detect unnatural language, we cannot study only syntactic aspects of the input text, as unnatural language can have a perfect structure. So the other aspect of the text we will study is its semantic. The objectives will not be to understand the meaning of the text. We will not tackle this still in progress research area, we will focus on coherency check.

4.1. *Statistical approach*

Generators can be tuned to respect word distribution. In this case we have to check consistency of the semantic of the text. We can do this at different levels. We can work inside blocks of text using methods like *cooccurences* matrix [LI 00]. The system is trained on a set of natural texts, collecting probability that each words has to appear with other words. Afterwards we can use the trained system to check, if words of the given text are not too far from each other.

Content generators can also work with existing texts, using some easily available sources like on-line books or websites. This is used for example in spamdexing to generate satellite pages to increase incoming links of a target web site. In this case, blocks of texts are sampled from several existing sources and concatenated. Keywords might be randomly inserted with links to the target web site. In this case we need to work with thematic relations between blocks of texts. We can use technics like those used in text segmenting ([HEA 97], [HEA 94]).

To achieve this we can compare *language models* of the different blocks (see [MIS 05] for the same method applied to detect blog spam comments). A language model is a statistical model representing the subject of a block of text : a probability distribution over strings, indicating the likelihood of observing these strings in a language. We compute language models of adjacent blocks of text and calculate the distance between them. In natural language, these distances tend to be small in most of the cases: the subject of a text does not change between each paragraph, unlike in some forms of unnatural languages.

4.2. *Symbolic approach*

Another approach to check semantic consistency is to use *wordnets* (see [STO 04]). A wordnet is a dictionary of semantic relations between words. It contains simple relations like synonymy or antonymy, but also more complex relations like specialisation, generalisation, "member of", "made of"... or contextual relations. Using these, we can check the consistency of text blocks and try to find full random text blocks or just inserted keywords.

For some elaborate unnatural text, this can be the most effective solution, but at the expense of a lot of manual work to produce the wordnet and probably coverage problems due to their limited size. In order to make wordnets more complete, but probably also less accurate, we need mixed approach using machine learning.

5. Perspectives

We have presented different methods to detect unnatural languages and the objectives of this thesis will be to evaluate them and see how they can be used and combinated for information retrieval applications.

These techniques have been already used in natural language processing. So, some tools already exist and we could try to reuse them. This must be done carefully, because most of them assume that their input text is natural language and have a real meaning. But unnatural languages does not always have meanings. In this case, some tools fail, but others will report unexpected results.

Like in all classification problems, a first step will be to prepare tests and training corpus to evaluate future works. We have to handle the classical difficulty of selecting carefully the type of text included in the different corpus.

First experiments will show us usability of the different methods proposed, and will dictate future works.

6. References

[GYO 05] GYONGYI Z., GARCIA-MOLINA H., "Web spam taxonomy", *In First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[HEA 94] HEARST M., "Multi-paragraph segmentation of expository text", *32nd. Annual Meeting of the Association for Computational Linguistics*, New Mexico State University, Las Cruces, New Mexico, 1994, p. 9–16.

[HEA 97] HEARST M. A., "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages", *Computational Lingustics*, vol. 23, num. 1, 1997, p. 33-64.

[LI 00] LI P., BURGESS C., LUND K., "The acquisition of word meaning through global lexical co-occurrences", *the 30th Child Language Research*, 2000, p. 167-178.

[MAN 99] MANNING C. D., SCHÜTZE H., *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, MA, 1999.

[MIS 05] MISHNE G., CARMEL D., LEMPEL R., "Blocking Blog Spam with Language Model Disagreement", *AIRWeb '05 - 1st International Workshop on Adversarial Information Retrieval on the Web, at WWW2005*, 2005.

[PAU 90] PAUL D. B., "Speech recognition using hidden markov models", *The Lincoln Laboratory Journal*, vol. 3, 1990, p. 41-62.

[STO 04] STOKES N., "Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking Domain", PhD thesis, Department of Computer Science, University College Dublin, 2004.

[ZIP 49] ZIPF G. K., *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Addison-Wesley, Cambridge, MA, 1949.