Extracting linguistic data from Usenet Newsgroups: troubles and challenges

Extracting linguistic data from Usenet Newsgroups: troubles and challenges

Claudio Russo, Phd Candindate
University of Turin, Italy
clrusso@unito.it

Abstract

This paper briefly explains why Usenet Newsgroups can be seen as source of interesting linguistic data. It also aims to illustrate the challenges that arise during the data extraction process and how they can be overcome resorting to pattern-matching functions.

1. Introduction

Usenet is one of the main sections that form the Internet (the other ones being the File Transfer Protocol, the World Wide Web and the e-mail). *Usenet* is very well known for hosting a huge number of communication platforms arranged by language and topic. Such platforms are called *Newsgroups*. Newsgroups' messages (or *posts*) can be visualized and downloaded with user-friendly client applications called *newsreader*.

Usenet Newsgroups' posts are particularly interesting as objects of linguistic analysis because of the lively, spontaneous language they are written in.[1]

2. Newsgroup post composition

A first, overall division of a Usenet Newsgroups post would split it into three main sections: metadata, the message text and signature lines.

Metadata provides the user with information about the post (such as author, newsgroup, publication date and time and so on) and it always appears within the post header. Its extraction can be elementary achieved using pattern-matching programming languages, since it always appears in a fixed structure. Scripts with anchors which are able to recognize the first field (i.e. the first sequence of alphanumeric characters between spaces) of the header's record will extract the metadata properly.

Signature lines are also easy to retrieve and extract: they appear at the bottom of the Newsgroup post, usually divided by the rest of the message by some sequence of even symbols ("--" most of the times).

An average post, in this particular domain, will be divided into quoted text and main message text. The challenges presented by both of these sections will be

1 For a deeper analysis of Usenet Newsgroups' language, please see Barbera & Marello (2008)

explored in the next paragraph.

3. Linguistic data and its challenges

During the process of data extraction, some elements deserve particular attention, since they can result potentially harmful in terms of linguistic analysis through corpus interrogation. Such items include:

- code strings (HTML and/or javascript code strings, attachments);
- web-art samples;
- spam;
- format errors in quoting structures;
- customized quoting marking;
- typing mistakes.

3.1 Code strings

Extracting relevant linguistic data from such strings might outcome into quite complex programming solutions with pattern-matching programming languages.

On one hand, simple tag sequences usually lead to straightforward solutions: relevant text strings embedded in simple HTML tags can be quickly selected and refined through substitution functions. On the other hand, complex tag structures embed the message text in such a way that its extraction would result too expensive, computationally speaking: in such posts, the unclear distinction between code strings and the message text forces the corpus builder either to create many *ad hoc* programming rules or to discard the post in order to dodge noisy data. Most of the times, the second solution results acceptable, since the deletion of some posts does not nullifies the thread's overall statistical validity.

Attachments deserve few separate lines because of their different encoding: programs, archives and pictures appear within the posts as extremely long, continuous sequences of alphanumeric characters. Pattern matching functions can easily recognize (and discard) attachments by setting up an acceptable field-length threshold combined with a short list of the code's most common sequences.

3.2 Web-art samples

Web-art samples are sequences of symbols, numbers, letters and space tabs shattered throughout the message window, in such a way that their overall perception outcomes in a picture. Most of the times, web-art samples carry no significant linguistic information and simple pattern-matching instructions result

sufficient to discard them.

3.3 Spam

Spam and cross-posting are two of the most threatening problems that may compromise the final corpus' representativeness. Spam posts can be tackled (and discarded) in the data extraction phase or during following refinement phases by checking each post's subject, message-id and newsgroup within the header.

Subject-based spam filters create an association chart in which all the thread's subjects
are temporarily stored and return their absolute frequency values. By establishing an acceptability threshold, any post which contains the subjects whose values scored above such threshold is discarded. Subjects starting with the sequence “Re:” or “RE:” must not be counted, in order to preserve any answer sent to a particular message.

Message-id-based filters check the unique alphanumeric sequence that identifies every single message and discard any message with the same id number in different newsgroups. This filter's acceptability threshold has to be extremely low to work properly: most of the times, a threshold of 2 identical message-ids results appropriate.

Cross-posting filters take into account the number of newsgroups that received the same message (i.e. messages with an identical id-number). They work exactly like message-id based filters (acceptability threshold aside, which is higher in cross-posting filters), but their use in the very first steps of data extraction makes the whole process lighter in terms of computational resources and expected duration.

3.4 Quoted text troubles

Quoted text patterns are designed to guide the Newsgroup user through previous conversation turns: each quoted text line begins with one or more closed angle bracket (>), each bracket representing one quotation level. In an error-free quotation structure any programming language based on pattern-matching functions can recognize the quotation level by the length of any quoted text's record first field.

Unfortunately, error-free quotation structures do not seem to be common: long quoted lines are often interrupted by the newsreader with a newline character and, when this happens, the newsreader does not insert any closed angle bracket at the beginning of the newly generated quoted line. Obviously, this situation does not cause any problem when a user is quoting another message for the first time, since all quoted lines will bear a quotation level of 1. Pattern-matching languages

come to a dead end in identifying quoting levels when this inconvenience occurs in more complex quoting structures (newsgroups' posts automated treatment showed up to the 25th level of quotation, but such levels are potentially infinite): at the current stage, guessing the missing quotation level recurring to pattern-matching functions seems too expensive in terms of computational resources, especially when a high number of different posts has to be automatically processed.

Customized quoting structures represent another (luckily uncommon) problem. Sometimes, users customize their newsreader's quoted text visualization preferring other symbols to the default “>”. When writing general, all-inclusive rules, this might be a problem: every time a line of the main message begins with a symbol, it would be classified as quoted text once the extraction function is called. It has to be said that such an uncommon situation can be ignored without of losing considerable amounts of relevant data. But, should the customization spread among the newsreader users, ignoring the problem may lead to statistical misrepresentations.

3.5 Typing mistakes

According to the linguistic register variation found across several threads, Usenet Newsgroups users seem to come from an extremely various range of sociolinguistic backgrounds. Such a wide register variation results precious in terms of data richness and representativeness, but it may also cause some difficulties during the extraction process.

Newsgroup posts can be downloaded from their server up to six months after their publication; Despite being available for quite a long period, they are still perceived as semi-permanent, a common feature among many Computer Mediated Communication platform. As a consequence, this perception makes the writer feel allowed to ignore some linguistic rules for written messages, assuming a more relaxed attitude towards morphology and syntax.[2]

Nevertheless, Newsgroups' users seem to pay particular attention the composing process: there are cases in which some users have been friendly scolded using word jokes for their spelling mistakes. Luckily, this spell-checking habit limits the occurrence of typing mistakes enough to allow their extraction without compromising any hypothetical research query. It looks probable that, at a certain point of the research project, an algorithm for the correction of the most common spelling mistakes will become necessary.

4. Conclusion

As spontaneous, computer-mediated, written linguistic products, Usenet

2 See Murray (1991), Scholz (2003) and Onesti (2007)

Newsgroups posts must be extracted ad arranged using algorithms that refines the raw data without compromising its natural linguistic features. Such features are, at least, desirable within an empirical basis that must respect some representativeness criteria, in order to allow the formulation of general linguistic principles.[3]

References

Barbera M., Corino E., Onesti C. (2007), Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup, in Barbera – Corino – Onesti (Eds.), *Corpora e linguistica in rete*, 25-88, Perugia, Guerra Edizioni

Barbera M., Marello C. (2008), Tra scritto-parlato, Umangssprache e comunicazione in rete: i corpora NUNC, in Accademia della Crusca (Eds.), *Studi di grammatica italiana – vol. 27*, 157-185

Murray, Denise E. (1991), The Composing Process for Computer Conversation, in *Written Communication – vol 8*, 35-55

Onesti, C. (2007), "Niusgrup"... si scrive così? Grafia in rete, in Barbera – Corino – Onesti (Eds.), *Corpora e linguistica in rete*, 253-270, Perugia, Guerra Edizioni

Scholz A. (2003), Comunicazione giovanile in rete: una mailing list italiana dedicata alla cultura hip-hop, in Rainer, Franz/Stein, Achim (Eds.), *I nuovi media come strumenti per la ricerca linguistica*, 117-139, Frankfurt am Main, Peter Lang

3 See Barbera, Corino, Onesti (2007) for full definition. Translation is mine.