Automatic Detection of Quotations in Multilingual News

Bruno Pouliquen, Ralf Steinberger, Clive Best
European Commission – Joint Research Centre
Via Enrico Fermi 1, 21020 Ispra (VA) Italy
{Bruno.Pouliquen, Ralf.Steinberger, Clive.Best}@jrc.it

Abstract

We present fully functional software that identifies direct speech quotations as part of its automatic analysis of more than 20,000 news articles per day in currently 11 languages. The system currently identifies over 2600 quotations per day, together with the person who made the quotation and – where applicable – the persons or organisations mentioned in the quotation. The most recent quotations from and about each person are listed on this person's dedicated information page, which is updated daily. As another component of the system also identifies variants of each name, the quotes can be assigned to the same person even if his or her name is spelled differently, allowing users to view all quotations from or about any of the currently 615,000 person names in the system's database in any of these languages. This automatic news analysis system is publicly accessible at http://press.jrc.it/NewsExplorer/.

Keywords

Quotation recognition, Named Entity Recognition, name variant merging, multilinguality.

1. Introduction

Many people and organisations are interested in finding quotations made by themselves or by other people in the world's media. The major interest groups looking for quotations are political analysts, company researchers and political actors. The motivation for the interest typically is the search for product feedback, for corporate image-relevant information, or for media feedback on political initiatives. We therefore developed an automatic tool that sieves through large quantities of media reports and extracts quotations plus the speakers and the persons referred to in the quotations. Due to the multilingual requirements in the European context, the developed quotation extraction tool had to be multilingual (it currently covers eleven languages). Due to this requirement, the applied methods needed to be simple and easy to extend to new languages.

In this paper, we first present related work (section 2) and the analysis data, i.e. the collection of media reports from which we extract quotations every day (3). We then give an overview of the method (4) and describe the details of the algorithm (5). In Sections 6 and 7, we present evaluation results and discuss them. Section 8 points to future work and draws conclusions.

2. Related Work

To our knowledge, there are only few online automatic systems that detect quotations by and about persons from text. Dimitrov et al. [6] developed a technique to resolve anaphora and applied it to quoted text (English only). There are a number of manually compiled websites that list famous or important quotations: *QuoteLand* (see [7]) allows to search for quotations by topic or author; *QuotationsPage* (see [8]) offers a large collection of historical quotes by known personalities; *WikiQuotes* (see [9]) is a compendium of several thousand user-collected important words in various languages, sometimes accompanied by their translation into English; *ThinkExist* (see [10]) is a large database of 300,000 English quotations, compiled over five years by more than 9,000 individuals. Most of these sites concentrate on historical quotations, and all of them are compiled manually. *DayLife* (see [11]) seems to detect recent quotes automatically in English language news. However technical details on how their system works are not known. Our own system, in comparison, automatically collects an average of over 2,600 quotations per day in eleven languages and is thus completely up-to-date. Currently, quotes are only listed on the relevant individual person pages of the JRC's NewsExplorer application (see [5]), but the plan is to make the collection searchable and display the most important quotes each day on a separate page (see Section 8 of this paper on future work).

3. Media Material

The JRC's *Europe Media Monitor* system (see [2]) gathers an average of 35,000 news article per day in 32 languages, by continuously monitoring about 1,000 public news sites from around the world for newly published information. The aggregated results are publicly accessible on the EMM-*NewsBrief* web site (http://press.jrc.it), which is updated every ten minutes.

The related EMM-*NewsExplorer* application (see [5]) clusters all articles gathered during the previous day by similarity in order to group all articles about the same subject or event. Each of these clusters is then further analysed to extract additional information, including the countries and geographical places mentioned and the references to persons and organisations. An average of 300 new person

names are automatically recognised every day together with the 'titles' they are associated with.

A database of all person names ever extracted by NewsExplorer is constantly updated fully automatically with the newly found information. This includes the information in which news clusters they appear, which other persons, countries and places they get mentioned with most frequently, which are the most common titles referring to them, etc. It is important to note that the whole process is automatic and that the displayed information is the result of statistics on extracted information from clusters of news.

This name repository is used to display a dedicated web page for each person, showing all the information the system was able to gather for this person (see Figure 1 as an example for information about Alexander Litvinenko).

In addition to clusters, countries and other associated persons we wanted to be able to detect automatically the quotations made by each of the persons in different languages. Moreover the quotations made by other persons about them was considered to be useful, too.

Figure 1. *Snapshot of part of the NewsExplorer page on the Russian spy Alexander Litvinenko, listing the automatically gathered name variants found in multilingual news and the most frequent titles and phrases that help to identify the name in running text. The example shows that different kind of information on Litvinenko (age, profession, nationality, death, etc.) was found in texts written in different languages.*

4. Method

As it was our aim to detect quotations in many different languages, we kept the linguistic input as simple as possible. We thus rely mainly on lexical patterns with character-level regular expressions, which are easily transposable to new languages.

As mentioned previously, our material consists of news articles in various languages (currently 32 in EMM). While we are aiming at detecting quotations in all these languages, we currently detect them in only eleven of them (Arabic, German, English, Spanish, French, Italian, Dutch, Portuguese, Romanian, Russian and Swedish).

The method used is quite simple: we look in the text of each article for quotation markers that are found close to reporting verbs (*say*, *declare* etc.) and known person names. For our purposes, known person names are those that have been found in at least five different NewsExplorer news clusters.

In most news articles, names found next to quotes are not full names consisting of first and last name. Common example types for quotations found in text are the following:

(1) Tony Blair said "We stand ready to support you in every way".
(2) "We stand ready to support you in every way," Blair said.
(3) Tony Blair visited Iraq… He said "We stand ready to support you in every way".
(4) Tony Blair visited Iraq… "We stand ready to support you in every way" the British Prime Minister said.

Our system currently only captures the first two types. Example (1) is not very common because the newspapers usually first talk about the context (*Tony Blair visiting Iraq*) and only then they introduce quotes.

Example (2) is more common and still easy to detect accurately. The issue here is that only the last name is mentioned and that we have to infer that the quote is by *British Prime Minister Tony Blair* even though there may be other persons with the name of 'Blair' in our database. We achieve this by first scanning the text for all occurrences of full names (consisting of first *and* last name), and by then assuming a coreference between the full name and the name part found.

In order to recognise the person doing the quoting in the third example, we would need to identify that the pronoun *he* refers to *Tony Blair*. We do not currently attempt to resolve such cases of anaphora because it would require additional language-specific effort and state-of-the-art anaphora resolution precision is relatively low. While [12] report up to 80 or 90% precision (below 80% with light-weight methods in [6]), the results for pronoun-drop languages like Spanish (see [13]), Italian or Korean only reach up to 74%. Anaphora resolution for pro-drop languages is less successful because subject pronouns are frequently omitted so that the gender of the subject is not made explicit in text. The following Italian quotation exemplifies this. We thus decided to ignore cases of pronoun use and to aim for higher precision, obviously to the detriment of the recall.

> *Luis Medina Cantalejo ha visto tutto. "La palla era altrove - __ racconta in un'intervista - e l'arbitro guardava in quella direzione"*
> where the subject of the verb *racconta* is not written (here indicated by __).

We do not currently try to identify the co-reference between 'British Prime Minister' and 'Tony Blair' in cases like (4), but have plans to do so. See the section on future work for details.

Our tool can rely on a highly populated database of names computed and updated daily as part of the NewsExplorer system. This database contains more than 615,000 names plus their variants, although we make only use of the 50,000 names (plus their 80.000 variants) that have been found in at least five different news clusters. The system is thus able to recognise any known name variants and to identify that they all relate to the same person. For instance, we have the following variants for the Uzbek president Islam Karimov: Islam Karimow (German), Islám Karímov (Spanish), Ислам Каримов (Russian), İslam Kerimov (Turkish), Islom Karimov (Swedish) and إسلام كريموف (Arabic).

5. Algorithm for quote recognition

We aim to detect all quotations accompanied by a named person as we cannot think of a use for quotations for which we do not know the name of the speaker. The system will recognise quotations only if it successfully detects three parts: the speaker name, a reporting verb and the quotation.

Our analysis of quotations in the news in various languages showed that many of the quotations are similar to the two examples below, i.e. the person making the quotation is either mentioned immediately before or after the quotation:

> "I don't think Congress ought to be running the war," Bush said yesterday.
>
> Mr. Wolfowitz said yesterday "I will accept any remedies".

What complicates matters is the use of anaphoric expressions instead of person names ('*he* said', 'added *the President*') and the fact that modifiers such as *yesterday* or *in a radio interview* may be found between the reporting verb and the quote. While we do not currently deal with anaphoric expressions at all, we do try to capture at least some modifiers.

5.1 Components for quotation recognition

Most quotations can be identified using a small number of rules. Our rules (Section 5.2) make use of the components described in paragraphs (A) to (F):

(A) quotation marker identification (quote-characters like ", ", «, » etc.)

(B) reporting verbs (e.g. *confirmed*, *says*, *declared* …)

(C) general modifiers, which can appear close to the verb (e.g. the adverb *yesterday*)

(D) determiners, which can appear between the verb and the person name (e.g. *the*)

(E) trigger-for-person (e.g. *British Prime Minister*)

(F) person name (e.g. *Tony Blair*)

(G) a list of matching rules (e.g. *name verb* [*adverb*] *quote-mark* QUOTE *quote-mark*)

We will now discuss these in detail.

(A) Quotation markers

In order to mark the quotation itself, we first identify and normalise the following quote-marks: ["] (two single apostrophes), [``] (two curly apostrophes), [,,] (two commas, used in some Dutch newspapers, [« /…/ »] (French quotes), [" /…/ "] (the English curly quotes), [<< /…/ >>] (two brackets), [" /…/ "] (double single-quotes), [' /…/'] (single quotes)

(B) Reporting verbs

They define a verb or any of its inflections that express that the string between quote-marks is a quotation. Without the presence of any of these verbs, we will not recognise the quotation. Examples are English *says*, *said*, *added*, *commented*, *sums up* and Italian *ha detto*, *dice*, *diceva*.

(C) General modifiers

These consist of quite generous lists of strings or regular expressions that are allowed before or after the verb. These strings are generally adverbs (*often*, *also*, *today*…), but there are also some compound expressions (*on television*, *last month*)[1]. We do not make use of external dictionaries, part-of-speech taggers or syntactic patterns. Instead, the list of modifiers has been derived empirically. To avoid listing all forms of verbs (*have said, might have said, would say*…), we also included the auxiliaries in this list of modifiers (in English: *has, have, had, would, might, could*, *do*, *did*, *does*).

(D) Determiners

In some cases, determiners can precede the name of a person. In our rules, they are allowed between the verb and the person name (English: *the*, French: *le, un, l'*, German: *der, die, seine*).

(E) Trigger-for-person

These patterns are usually titles of persons (*Dr., Prime Minister, French President*…). However, we prefer to call them trigger-for-person because they could be more gen-

[1] The Spanish configuration includes the following regular expression (*por la* |*en la* |*a la* |*en*)(*mañana*|*tarde*) recognising *por la mañana* or *a la tarde*.

In French: *pour sa part* and even the days of the week (*lundi*, *mardi*…) as it is quite common to say in French: *"…" a dit lundi Jacques Chirac.*

eral expressions referring to nationality (e.g. *the Iranian*), age (*57-year-old*) or other. In a random set of 240 English quotations, we found that in nine cases (3.75%) the title of the person was found before the person name. This low number is presumably due to the fact that the titles are used when the person is first introduced while quotes are usually mentioned further down in the article.

For the detection of names in NewsExplorer, we built (semi-automatically) an extensive list of such trigger words. In English, the list currently comprises more than 1,000 items. Recognition patterns also allow for combinations of several of them (e.g. *young Spanish Ambassador*).

(F) Person name

The most important person names are automatically detected as part of the daily process for NewsExplorer (see specifically [1]). About 50,000 person names and their variants are compiled into an automaton, which is updated every day. The person names are then marked up in each article. In order to resolve the name part co-reference resolution, we then look up in text the uppercased words that are also part of a full name found elsewhere in the text. This method can identify 'Tony Blair' as the author even if only the last name of the author is used in the text (e.g. [*Tony Blair*] *visited Iraq yesterday. … "I reiterate our determination to stand four-square behind you" said* [*Blair*]).

5.2 Matching rules

In order to write the quotation matching rules, we first had to carry out a survey of the various ways to express a quotation across languages. We found three generic rules and a number of additional language-specific rules.

The three generic rules are:

(1) *quote-mark* QUOTE *quote-mark* [,] *verb* [*modifier*] [*determiner*] [*title*] *name*
e.g. *"blah blah", said again the journalist John Smith.*

(2) *name* [, *up to 60 characters* ,] *verb* [:|that] *quote-mark* QUOTE *quote-mark*
e.g. *John Smith, supporting AFG, said: "blah blah".*

(3) *quote-mark* QUOTE *quote-mark* [; or ,] [*title*] *name* [*modifier*] *verb*
e.g. *"blah blah", Mr John Smith said.*

The following format was found only in Italian and Russian articles:

(4) *quote-mark QUOTE1 - [modifier] verb name - QUOTE2 quote-mark*
e.g. "Ciampi – ha detto Berlusconi – ha favorito la sinistra perché era un uomo della sinistra"
where the author (here Berlusconi) and the reporting verb (*said*) is included *inside* the quotation marks, marked by hyphens.

The Swedish writing convention for quotations includes sentences beginning with one or two hyphens "--":

(5) -- QUOTE, *verb* [*adverb*] [*title*] *name*
e.g. -- Vi försökte uppmuntra samverkan, säger Urban Lundmark.

A specifically Arabic pattern is to mention the verb *before* the person name. We therefore introduced the rule:

(6) *verb* [*title*] *name* [*modifier*] *quote-mark* QUOTE *quote-mark*
[and said minister of justice Saddam Hussein to Israel radio "we don't .."]
وقال وزير العدل صدام حسين لإذاعة إسرائيل
" إننا نحمل عباس المسؤولية النهائية عما يحدث".

6. Evaluation of quotation recognition

Users can consult the quotations of each person in NewsExplorer. The process gathers an average of 2,665 quotes per day (1647 of which are found in 7000 English articles every day). As of June 2007, we have a repository of about 1,500,000 quotes, gathered during 2 years of analysis. This repository is not currently fully exploited apart from displaying quotations of/about a person as part of the NewsExplorer's person pages. From an application-oriented point of view, this works rather well: For many persons, NewsExplorer displays recent quotes from or about the person in many different languages.

In order to evaluate the *Recall* of the quotation recognition system, we searched a random collection of news articles (documents dated 12 July 2007) for any of the quotation markers mentioned in Section 5 and carried out a manual evaluation for 55 of the quotations found. We found that a surprisingly high number of 42 examples (76%) were quotations our system does not actually try to identify. Most of these 42 quotations were by persons whose name was not mentioned at all in the article (e.g. *the officer / their neighbour*). The remaining ones were by persons that are not part of our *known persons* (i.e. persons that have been found in at least five different news clusters over the past few years). For the remaining 13 cases, i.e. those that do fall inside our mandate and that we do try to identify, seven were correct while 6 had not been found, corresponding to a Recall of 54%. However, all of the six quotations that had been missed at document level had been found in other articles, so that the Recall *within the news collection* was in fact 100%. This finding confirms that we should aim for precision rather than recall because of the data redundancy in the EMM news collection.

The reasons why the seven investigated quotations had not been found are the following (multiple counting is possible): One quote was not identified because the speaker was only represented by a pronoun (*he*). In one case, our rules did not match because the verb form was missing (*telling* – this has now been added to the rule). In one case, the speaker's name was badly tokenised, leading to non-recognition: For UN Secretary General *Ban Ki-moon*, our

system identified *Ban Ki* as the name and the remaining string *moon* stopped the rule from recognising the quotations (The tokenisation bug has now been fixed). The largest source of errors, however, were unknown modifiers (three cases, including *in a short statement*, *with relief*), leading again to non-recognition. As not all possible modifiers can be captured with our simplistic rules, such cases could only be solved by making use of a full morpho-syntactic analysis of the sentence. The only erroneous quotation recognition was an incomplete quote: Only the first part of the quote was found while the second part of the quote (continued after an interruption) was missed. This case lowered the overall *Precision* for the English language evaluation to 87.5% (7/8).

In order to evaluate the *Precision for multilingual quotation detection*, we carried out a second, mixed-language evaluation: Out of the 1,500 quotations of a given day (17/12/2006), we randomly selected 120 in 10 languages (discarding two quotations of the same person in the same language). The test set contained 1 Arabic, 10 German, 41 English, 22 Spanish, 4 French, 14 Italian, 3 Dutch, 16 Portuguese, 3 Russian and 6 Swedish texts. An expert read each article where the quotation was detected and judged the quality as "correct", "incomplete" or "wrongly assigned". An *incomplete* quotation is when only part of the full quotation was found, i.e. the system detects the first part of the quotation, but misses its continuation, as in the example:

> "*I'm really happy for Fabio," Materazzi told the Apcom news agency Friday.* "*I feel part of this distinction because I think that all the Azzurri helped a great champion like Cannavaro win an important prize*".

In this case, only "*I'm really happy for Fabio,*" was detected by the system, while the continuation was missed. A *wrongly assigned* quotation is one where the quotation was uttered by another person than the one identified by the system. An example for such a wrongly assigned quotation is the following:

> *Le porte-parole du Haut représentant de l'UE pour la politique extérieur Javier Solana a jugé "condamnable" le saccage du terminal de Rafah…*[the spokesperson of the EU High Representative for external policies Javier Solana judged "reprehensible" the devastation of the Rafah terminal].

The system detected "*condamnable*" as a quotation, but attributed the authorship to Javier Solana, while it should have been attributed to his spokesperson.

The mixed-language evaluation yielded the following results, by category: Correct: 81.7%, incomplete: 17.5%, wrongly assigned: 0.8% (one document).

7. Discussion of the results

Taking into account the simplicity of the approach, we consider the overall results to be rather good. The Precision is rather high, and the relatively low Recall at document level is often compensated by the data redundancy, i.e. the same quotation will frequently be found in another news article.

Obvious restrictions of the approach are the following:

- There is no co-reference resolution for pronouns and for titles (trigger-for-person);
- There is no recognition of unknown modifiers that separate the reporting verb and the quotation (no parsers are used to recognise adverbials in the shape of adverbs, noun phrases such as *with relief* and prepositional phrases such as *in a short statement*).
- Quotes in genitive constructions are currently assigned to the wrong person (In *"…" said Blair's spokesperson*, Blair would be identified as the author of the quote).

However, the simplicity of the system also has important advantages:

- The process is fast and can detect a high number of quotations in only a few seconds.
- Multilinguality is not an obstacle: NewsExplorer is currently handling eleven languages for quote recognition and gathers quotations of the same person in many news articles from around the world.
- The system is fully automatic. It currently runs every morning and adds new quotations of the last day to every person page.
- Time and source of the quotation are identified and displayed. The user can thus always read the full article (if it is still available on the original website) to verify the correctness of the quotation.

8. Future work / Conclusion

We would like to improve the accuracy of the recognition. As the evaluation showed, a full morpho-syntactic analysis of the sentences containing quotations would be beneficial, especially to deal with the wide range of adverbials that cannot all be listed as part of our simplistic rules. The cost for a full sentence analysis, however, would be that the tool would be less easily extendable to new languages because a different parser would be required for each language.

We are aware that pronoun co-reference resolution would be an important step towards increasing the recall of the system, although the error rate in anaphora resolution might lead to wrongly assigned quotations, which we want to avoid as much as possible. Instead, we may want to focus on the co-reference between titles (e.g. *Spanish Prime Minister*) and names (*José Zapatero*), by making use of the wealth of information in NewsExplorer on person names

and their frequently attributed titles. This would help to attribute quotations correctly in sentences like the following:

> *[José Zapatero] visited France on Monday. "We are friends" said the* [*Spanish Prime Minister]*.

It might not be too difficult to link multi-part quotes (*"Yes we do," declared John "we will win"*), using relatively simple patterns. We should investigate this.

Regarding the usage of the output of the system in NewsExplorer, we would like to offer a separate page showing the most important quotations of the day. This would require finding a criterion to rank the quotations. An idea would be to make use of multilinguality to show first the quotations by persons having made most of the quotes of the day across all languages. In this context, we also plan to develop an interface allowing users to search quotations by name or using free-text search.

We have started experimenting with detecting the sentiment of quotations and to classify them into positive and negative statements. News analysts may be rather interested in knowing the attitude of public figures towards certain themes or persons.

As part of a larger effort to extract specific relations between persons (e.g. Tanev 2007, Pouliquen et al. 2007), we plan to build a *quotation network*. The idea here is to identify a social network based on who makes reference to whom in their quotations (see Figure 2 and the prototype application at http://langtech.jrc.it/picNews.html).

Our system is now fully functional and identifies about 2,600 quotations per day in eleven languages. The quotations from and about a person are publicly accessible at the site http://press.jrc.it/NewsExplorer/.

The NewsExplorer website is very popular (getting up to 1,200,000 hits per day), among other things because it compiles information about over 615,000 persons. The quotations (from the person, or about a person or organisation) contributes to this success. The multilingual aspect presumably is a determining feature, as well. Future developments will make the quotations more visible to the end-user.

9. Acknowledgement

We thank the following persons for their help in developing language-specific resources for quotation recognition: Anna Widiger (Russian, German), Camelia Ignat (Romanian), Wajdi Zaghouani (Arabic), Bart Wittebrood & Tom de Groeve (Dutch) and Ann-Charlotte Forslund & Patrik Hoglund (Swedish). We thank Jenya Belyaeva for the evaluation of the results. Our special thanks go to the entire EMM team and especially to Flavio Fuart, who helped put the results online.

10. References

[1] Best, C., van der Goot, E., Blackler, K., Garcia, T., Horby, D. (2005). Europe Media Monitor – System Description. Report No. EUR 22173 EN.

[2] DayLife (2007) http://www.daylife.com/, last visited 12/02/2007

[3] Dimitrov, M., Bontcheva, K., Cunningham, H., Maynard, D., (2004) A Light-weight Approach to Coreference Resolution for Named Entities in Text, Anaphora Processing: Linguistic, Cognitive and Computational Modelling, Antonio Branco, Tony McEnery and Ruslan Mitkov (editors)

[4] Mitkov, Ruslan (2002). Anaphora Resolution. Longman.

[5] Pouliquen, B., Steinberger, R., Ignat, C., Temnikova, I., Widiger, A., Zaghouani, W., Žižka J. (2005). Multilingual person name recognition and transliteration. Journal CORELA - Cognition, Représentation, Langage. Numéros spéciaux, Le traitement lexicographique des noms propres..

[6] Pouliquen Bruno, Ralf Steinberger & Jenya Belyaeva (Submitted). *Multilingual multi-document continuously-updated social networks*. Workshop *Multi-source, multilingual Information Extraction and Summarization* at RANLP'2007.

[7] QuoteLand (2006). http://www.quoteland.com/, last visited on 20.12.2006.

[8] QuotationsPage (2006). http://www.quotationspage.com/qotd.html, last visited on 20.12.2006.

[9] Steinberger, R., Pouliquen, B., Ignat, C., (2005) Navigating multilingual news collections using automatically extracted information. Journal of Computing and Information Technology - CIT 13, 2005, 4, 257-264.

[10] Tanev Hristo (2007). *Unsupervised Learning of Social Networks from a Multiple-Source News Corpus*. Proceedings of RANLP'2007.

[11] ThinkExist (2006). http://en.thinkexist.com/, last visited on 20.12.2006.

[12] WikiQuotes (2006). http://en.wikiquote.org/, last visited on 20.12.2006.