Networking Readers: Using Semantic and Geographical Links to Enhance e-Books Reading Experience

Dan Cristea[1,2(✉)], Ionuț Pistol[1], Daniela Gîfu[1(✉)], and Daniel Anechitei[1]

[1] Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi, Iaşi, Romania
{dcristea,ipistol,daniela.gifu,daniel.anechitei}@info.uaic.ro
[2] Institute for Theoretical Computer Science, Romanian Academy - Iaşi branch, Iaşi, Romania

Abstract. This paper describes how a system currently developed can be used to connect readers of enhanced e-books both to each other, to web resources and to real world locations and events. A set of Natural Language Processing resources are used to annotate relevant e-books and a framework is developed using the original text and the annotated metadata to detect and display semantic connections within the text and from text to relevant web data. This system can be further enhanced to detect connections between users using common reading interests and habits, their location in relation to locations found in text, and their reading and real world localisation history. Users could also be able to share collected data (text and web references, video and audio recordings, other interested readers) improving an individual reader experience and helping to establish a community around a particular e-book or a real life location with literary significance.

Keywords: Social networks · Semantic links · NLP · Geographic data · e-Books

1 Introduction

Building communities of users around technologies can enhance the ways users' benefits from the implemented functionalities and can increase both the frequency and the duration of usage of said technology. E-books have become a common way of experiencing written content, the most significant contributing factor being the increasing usage of mobile devices both for communication and for access to documents.

This paper aims to show how a currently developed system designed to add semantic links within texts and between text and outside resources (web pages and maps) can be further developed to build communities of users around particular E-books using discovered links (mostly of geographical and social nature).

Employing text mining techniques to improve social networks is not a novel idea. Ever since the emergence of large networks of user contributed data (mostly text), researchers have described methods of using that text to extract additional data about the network and its users [1, 2]. Two areas in which Natural Language Processing technologies have been particularly successful in enhancing a social network user experience have been e-learning [3, 4] and consumer reviews [5, 6]. In both cases, links between users are established based on their consumed and produced data and in both cases this is shown to significantly increase users satisfaction and user retention.

N.T. Nguyen et al. (Eds.): ICCCI 2016, Part II, LNAI 9876, pp. 59–67, 2016.
DOI: 10.1007/978-3-319-45246-3_6

The novelty of our approach is in applying these techniques in a richer text environment, using heavily annotated texts (with both syntactic and semantic metadata) especially including real-world geographical references. Quite often a reader feels the need to supplement the knowledge on certain places or notorious people that are mentioned in the book she/he reads by searching the web or visualising places on geographical atlases (in digital or classical printed forms). MappingBooks (Fig. 1) is a technology that facilitates these searches by pre-computing links outside the book text itself.

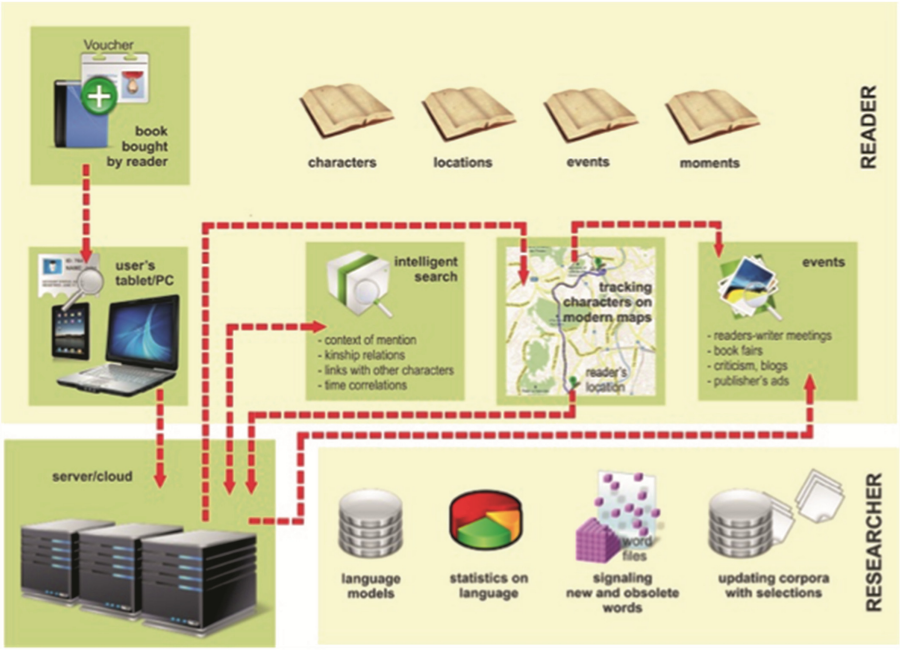

Fig. 1. MappingBooks in a bird's eye view

Installed on mobile devices, and accessible by readers on the basis of subscriptions, the technology will enrich the texts with contextual links intended to enhance the reading satisfaction, offering supplementary and interactive guide beyond the text itself and making the reading entertaining and pleasant.

When comparing the proposed functionalities with those offered by other social communities built around reading hobbits (such as Goodreads[1]), our approach offers a closer connection to the actual content of the books, basing most detected connections between users on the existing metadata previously added to the books, as described in Sect. 4 of this paper.

The paper has the following structure. In Sect. 2 we briefly describe the way documents are processed in MappingBooks, detailing the annotations the technology adds

[1] https://www.goodreads.com/.

to the raw text. Section 3 shows what types of semantic relations are currently added to texts and what further efforts could be made in this regard. Section 4 describes how users can be connected using both the viewed text content and their activity. Finally, Sect. 5 summarises the proposed development efforts and shows how they could further benefit other systems.

2 Text Mining in MappingBooks

A MappedBook, the technological output of the MappingBooks (MB) ongoing project is an online connected book that facilitates creation of social networks based on reading preferences. Mentions of locations and other entity names, which the book contain, are automatically identified and put in correlation with geographical artifacts (like maps, coordinates, layouts) and the web. As such, the user (considered to be reading the book while connected to the application) will be directed, at appropriate moments, towards significant events happening in the real world in locations mentioned in the book, which are reflected in the virtual world. Moreover, sensible to the instantaneous location of the user, the system can locate her/his position in connection with geo-mentions contained in the book, creating thus a more intimate relationship between the text and its readers.

Hypermaps [7] composed of background base layers with continuous cover (mainly represented by raster data) or overlay layers, as discrete, raster and vector data, will be associated to geonames in text, thus offering additional multimedia information, mainly through pop-up windows and hyperlinks.

To support these functionalities, the base text is automatically processed as follows. First, the base text is extracted from the original document (usually a PDF file which includes images or other graphical artifacts). This step is performed by using first the iText library, then by manually correcting the resulted text (fixing diacritics and hyphen segmented words, removing page numbers, image captions and other additional elements). All text extraction errors are also fixed, the quality of the resulted text being extremely important for increasing the accuracy of the further automated processing.

The corrected text is then passed through a series of annotators: POS tagger [8], NP-Chunking [9], NER (Name Entity Recognizer) [10] and RARE (Robust Anaphora Resolution Engine) [11]. The result is a stand-off annotated XML document which is then used as input for the following semantic processing.

3 Linking Entities in and Out of Text

Entities are uniquely identifiable physical/abstract things. Entity detection (i.e. finding a string denoting an entity) and understanding (linking the string to an entry in a knowledge repository where the entity is described) takes us a long way towards understanding the text itself, and provides additional knowledge to the reader. The detection of geographical entities belonging to 15 classes is done by a Named Entity Recogniser [13].

External repositories, such as Wikipedia and Open Linked Data, which combine numerous, rather stable sources of knowledge, are used to link the mentions outside the book. This process is illustrated in Figs. 2 and 3.

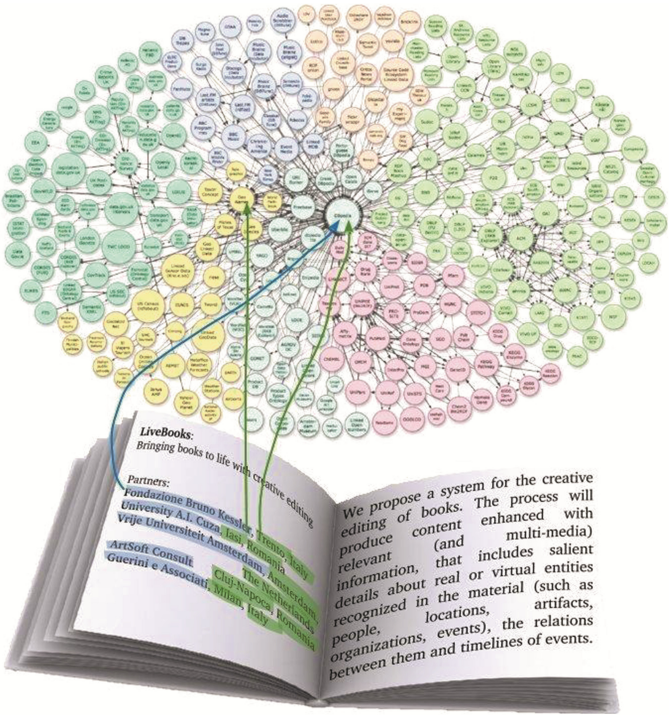

Fig. 2. Linking entities to Open Linked Data.

Our aim is not only to find additional information, but interesting additional information. The notion of interestingness, used in the fields of knowledge discovery and text summarization, contrasts the notion of importance. Important things may be common knowledge, while something interesting is idiosyncratic, specific to a certain reader, unexpected and previously unknown. The following reader profiling sources are used in our project: age, sex, profession, nationality, city and region address, musical, reading and other cultural preferences, hobbies, etc. Profiling the user means filling in a vector of characteristics (by using different sources agreed by the user to be connected to the application, the least liked being direct acquisition at sign up). Then, for each piece of online linked information a similar vector is filled in, by using contextual sources (as given by headers, origin of sites, etc. and using bag of words and tf*idf measures). Then, vectors of acquired pieces are matched against the user's vector and only the best ranked are retained.

Entity mentions in a text are said to be coreferent when they refer to the same entity. Mentions may take different syntactical forms in a text, but the most common ones are noun phrases built around proper nouns (named entities), common nouns or pronouns. While reading a book, the reader continuously deciphers coreferential mentions, most of the time without a conscious effort. Coreferential (and in general, anaphoric) relations give a text cohesiveness, allowing a reader to connect entities between them and connect events through their participating entities, to build the picture the author intended. The set

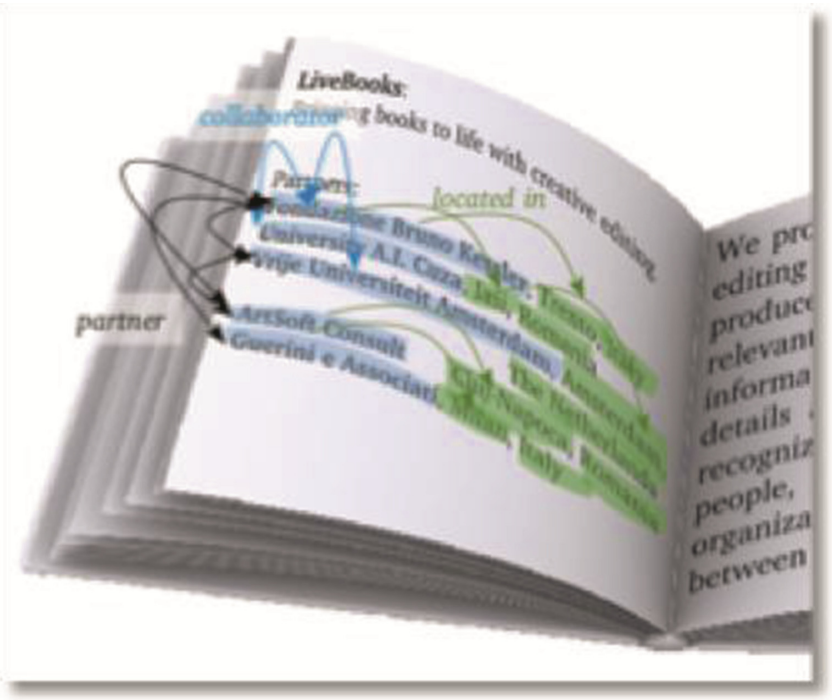

Fig. 3. Entities and the links between them.

of relations that link mentions referring to the same entity over a whole text can take the form of a chain or a tree, with links always directed from the current mention, called anaphor, to coreferent mentions that appeared before them in the text – antecedents [12]. Thus, to each entity that appears in a book corresponds a chain or a tree of coreferential links, with the root usually anchored in the first mention of the corresponding entity. We call this structure the coreference chain. Coreference chains are important for building and synthesizing information from a book. Extracted information about the entity (e.g. images, URL to Wikipedia or Open Linked Data) will be shared by all nodes in the chain. It is also important to distinguish shared information from information particular to a specific mention. Temporal or geographical coordinates, for example, may not be shared– the same entities may appear in different locations at different times. Such information must remain linked only to the specific mention to which it applies, based on the context.

It makes explicit the participants in events by detecting the entities to which they refer, thus preparing the text for the event analysis phase. It is also crucial in establishing connections between events that share participants, and it may help detect relationships between entities [13]. Relations between different types of entities may include: people are located in specific places, events occur at specific points in time, mutual positioning of locations in space, spatial distances and directions, etc., as exemplified in Fig. 2. In MB there are identified 17 types of semantic relation between entities [14], adding to the outside text relations mentioned above.

4 Linking Users

As the system will gain in popularity, the possibility to connect the community of users based on a diversity of sub-group preferences becomes interesting. Users have the option to declare their willingness to be included in such a community, in which case shared users' open data and preferences (education, readings, sensibilities, music and travelling preferences) as well as instantaneous contexts (as is the immediate location, history of past travels or the intention to start a journey) can be used as selection criteria.

As described in the previous sections, the annotation automatically added to the text, depending on the content, could be extremely rich. Among them there should be entities with geographical real-world significance, institutions and notorious people, as well as links between these entities both inside and outside the text. For registered users the system keeps personal data, such as identification information, collections of MB books subscripted for, instantaneous locations of devices and histories of users' trajectories (previous locations). Part of this information, the stable data, can be obtained, with users' accepts, from social networks (Facebook, Twitter, LinkedIn, Google +) where they are members, other data are typed in at registration, while the volatile, quickly changing, data, such as location and journeys is collected during daily or sporadic use.

Adopting the trendy typology of accessibility of data from current social media networks, here too the user's data can be: public, restricted to classes of friends or for personal use only. Following, is an incomplete list of types of connections that can be established between users, based on visible data:

- If a user has declared the information on books "subscripted for" as visible, then the system can form the user's community **current co-readers of B**. A user is in this community if MB actively changes information about B with her/him. Users of this community would thus be able to share hot impressions on the reading B, supplement their reading lists with similar suggestions; invite people to visit their personal cultural forums, etc.
- The above community can be enlarged if the necessity to read the book B at the very current moment is removed. A user is in the **co-readers of B** community, even if not online, if MB knows about her/him to have been changing data about this book with her/him at any time now or in the past.
- If a user has declared as visible the information on "instantaneous location", then the system can form the current **co-proximity of L** community. A user is in this community if MB knows about her/him that is currently online and is physically located in that location or in proximity of it. This type of information can create very interesting links between readers, based on personal impressions, photos, shared real-time descriptions.
- Again, the above condition can be relaxed if the demand to be in the location L at this very moment is removed. A user is in the **co-proximity of L** community, even if she/he is not currently online, if MB knows about her/him that has been at any time, now or in the past, in that location or in proximity of it.
- If the location is enlarged to include points assigned to a journey, communities of people sharing a route (track) can be formed. The community represents a

generalisation of the MB **co-proximity of L** community, known as **co-track of T**. A user is in this community, even if she/he is not currently online, if MB knows about her/him that has been at any time, now or in the past, in any of the points belonging to track T or in proximity of it. Also as above, the subcommunity of people sharing the same track at the current moment, **current co-track of T**, can be formed.

- Any combination of the above species of communities can be dynamically formed on request by the user. As such, a user can contact very constrained communities such as current **co-readers of B** AND **current co-proximity of L** down to much less constrained, such as **co-readers of B** AND **co-track of T**. Thus, links can be established between people that have common reading preferences and have pursuit similar travelling experiences on places inspired by these lectures.

All described connections are unilaterally established from the initiating user (the active user). They dont require any action/confirmation from the other users involved, aș they dont make available to the original user any sensitive data (such as personal info, search/reading history, travel history). All user data shared is only relevant in the local context it is made available.

Since the original MappingBooks system is designed as a client-server application, considerations have to be made with respect to the usage of the limited memory and processing resources available on the device, the client. The features described in this section are under development, but we believe that adding social networking capabilities is feasible, without a major impact on processing or memory overload of the client-server application. This is because all relevant data on users should be kept on the server, as specific data paired with the users' IDs. Then, after receiving from a user the type of community asked for, they are formed by filtering conditions expressed on a common database and only results will be communicated back to users. Actually, following classic algorithms used in artificial intelligence that minimise the matches between data and patterns [15], a fixed but large number of predefined communities can be updated permanently and their retrieval made instant.

5 Conclusions

At the present moment MappingBooks offers a basic mobile app serving as proof-of-concept for part of the functionalities described above. As an example of relevant types of e-books the app currently offers an annotated geography manual (described in [13, 14]). The included annotations are initially added automatically (at the surface morpho-syntactic level), and further enhanced with manual annotation of entities and semantic relations.

The proposed functionalities are well within the technological capabilities of current mobile devices. The MappingBooks system and its connection facilities addresses a diversity of possible users: from the passionate readers, people enjoying to read books everywhere, to occasional readers, those reading only during travelling or in vacations, from youngsters, school children and students to retired people, from adventures, those often on route to people travelling only in their minds, who have never stepped out of their town. Linking entities identified in text to Open Linked Data will project the book

into a huge semantic space, which may contain snippets of information that would make useful additions to the book content. Moreover, finding people with similar reading preferences and easily establishing contacts with them will be enjoyable and rewarding. Novels, biographies, books with historical or geographical subjects, class manuals and travel guides are only some examples of styles that are lend of being transposed in the MB technology.

Publishing houses could be the principal beneficiaries of the MappingBooks technology, as it could generate increased book sales over time if correctly mastered from the point of view of business models. A system of bonuses may also bring advantages to publishing houses in partnership with local administration or tourism agencies, as for instance, one that would challenge the readers to hit as many of the books places, visiting all locations or getting through all mentioned routes. A way in which providers of tourism services could be informed on the user's travel interests and particular locations and routes associated with a popular new book can also be easily imagined.

With respect to forming communities, the MB interface can be extended to include lists of "achievements" in connection with a text. Thus, users can be automatically upgraded for "connoisseurs" levels with respects to, for instance, tourist objectives mentioned in text.

It is easy to imagine other ways to form communities rooted in lectures, as, for instance, selections that intersect common readings and attended places with levels of friendship reported by other social media, like Facebook or Twitter. Events and entities mentioned in a book can be associated with a real-world location and a particular time of the year (or of the day). If a history is available for users' locations, a reader can identify users who visited that location and/or witnessed that event at the relevant time of the year/day.

However, there remains the problem of finding a few needles in numerous haystacks, and putting them together into a coherent whole, or otherwise the reader will soon be suffocated by the amount of useless information made available. Even if the text may contain clues that can be used as constraints about what relevant information is (such as the time frame where the entity is mentioned, the location and the general context), we are not yet totally clear about the right way to filter the linked information.

Acknowledgements. The work reported in this paper was achieved with the support of the PN-II-PT-PCCA-2013-4-1878 Partnership PCCA 2013 grant "MappingBooks - Intră în carte!", having as partners UAIC, SIVECO and „Ștefan Cel Mare" University of Suceava. We address our thanks to Vivi Năstase for relevant ideas and realisation of Figs. 2 and 3.

References

1. Aggarwal, C.: Text mining in social networks. In: Aggarwal, C. (ed.) Social Network Data Analytics, 2nd edn, pp. 353–374. Springer, Heidelberg (2011)
2. Irfana, R., Kinga, C.K., Gragesa D., Ewena S., Khana, S.U., Madania, S.A., Kolodzieja, J., Wanga, L., Chena, D,, Rayesa, A, Tziritasa, N., Xua, C., Zomayaa, A.Y., Alzahrania, A.S., and L, H.: A Survey on Text Mining in Social Networks. The Knowledge Engineering Review. Cambridge University Press, Cambridge, pp. 1–24 (2015)

3. Romero, C., Ventura, S.: Data mining in education. Wiley Interdisc. Rev. Data Min. Knowl. Disc. **3**(1), 12–27 (2013)
4. Hung, J.L., Zhang, K.: Examining mobile learning trends 2003–2008: a categorical meta-trend analysis using text mining techniques. J. Comput. High. Educ. **24**(1), 1–17 (2012)
5. Cambria, E., Schuller, B., Xia, Y., Havasi, C.: New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. **2**, 15–21 (2013)
6. Netzer, O., Feldman, R., Goldenberg, J., Fresko, M.: Mine your own business: market-structure surveillance through text mining. Mark. Sci. **31**(3), 521–543 (2012)
7. Kraak, M.-J., Rico, V.D.: Principles of hypermaps. Comput. Geosci. **23**(4), 457–464 (1997)
8. Simionescu, R.: UAIC Romanian Part of Speech Tagger, resource on nlptools.info.uaic.ro, "Alexandru Ioan Cuza" University of Iași (2011)
9. Simionescu, R.: Romanian deep noun phrase chunking using graphical grammar studio. In: Moruz, M.A., Cristea, D., Tufiș, D., Iftene, A., Teodorescu, H.N. (eds.) Proceedings of the 8th International Conference Linguistic Resources and Tools for Processing of the Romanian Language, pp. 135–143 (2012)
10. Gîfu, D., Vasilache, G.: A language independent named entity recognition system. In: Colhon, M., Iftene, A., Barbu Mititelu, V., Cristea, D., Tufiş, D. (eds.) Proceedings of ConsILR-2014, pp. 181–188, "Alexandru Ioan Cuza" University Publishing House, Iaşi (2014)
11. Ignat, E.: RARE-UAIC (Robust Anaphora Resolution Engine), open-resource on META-SHARE, "Alexandru Ioan Cuza" University of Iași (2011)
12. Cristea, D., Postolache, O.: Anaphora resolution: framework, creation of resources, and evaluation. In: Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages, FASSBL-2006, 18–20 October, Sofia, Bulgaria (2006)
13. Cristea, D., Gîfu, D., Pistol, I., Sfirnaciuc, D., Niculita, M.: A mixed approach in recognising geographical entities in texts. In: Trandabat, D., Gîfu, D. (eds.) EUROLAN 2015. CCIS, vol. 588, pp. 49–63. Springer, Heidelberg (2016). doi:10.1007/978-3-319-32942-0_4
14. Gîfu, D., Pistol, I., Cristea, D.: Annotation conventions for geographical relations. In: Proceedings of the 11th International Conference Linguistic Resources and Tools for Processing the Romanian Language, ConsILR-2015, pp. 67–78 (2015)
15. Waterman, D.A., Hayes-Roth, F. (eds.): Pattern-Directed Inference Systems. Academic press, New York (2014)