Knowledge-Based Systems 108 (2016) 155–167

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Merging open knowledge extracted from text with MERGILO

Misael Mongiovì[a,*], Diego Reforgiato Recupero[a,c], Aldo Gangemi[a,b], Valentina Presutti[a], Sergio Consoli[a,d]

[a] Semantic Technology Lab, ISTC-CNR, Rome and Catania, Italy
[b] Paris Nord University, Sorbonne Citè, CNRS UMR7030, France
[c] University of Cagliari, Via Ospedale 72, Cagliari 09124, Italy
[d] Philips Research, Data Science Group, High Tech Campus 36, 5656AE Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 15 November 2015
Revised 25 April 2016
Accepted 8 May 2016
Available online 10 May 2016

Keywords:
Knowledge reconciliation
Coreference resolution
Knowledge base integration
Graph alignment

ABSTRACT

This paper presents MERGILO, a method for reconciling knowledge extracted from multiple natural language sources, and for delivering it as a knowledge graph. The underlying problem is relevant in many application scenarios requiring the creation and dynamic evolution of a knowledge base, e.g. automatic news summarization, human–robot dialoguing, etc. After providing a formal definition of the problem, we propose our holistic approach to handle natural language input – typically independent texts as in news from different sources – and we output a knowledge graph representing their reconciled knowledge. MERGILO is evaluated on its ability to identify corresponding entities and events across documents against a manually annotated corpus of news, showing promising results.

1. Introduction

This paper focuses on the problem of acquiring knowledge from multiple *natural language* (NL) sources and *reconciling* it in an *integrated formal representation*. This problem, referred to as *knowledge reconciliation*, is relevant in most application scenarios that require to create and evolve a knowledge base from multiple and dynamic NL sources, for example: (1) building an integrated knowledge view, e.g., a summary, about a specific event, e.g., the Opening of 2012 London Summer Olympics, by acquiring knowledge from different newspapers [1]; (2) supporting human–machine dialogue in the context of assistive robotics by collecting a patient's personal memories, which are provided through NL inputs over time. Let us consider the following news from two different sources:

"Tony awards: "Fun Home" and "Curious incident" big winners."

and

"On Broadway's biggest night "Fun Home" wins Tony award for Best Musical"

In an ideal scenario, the goal is to automatically produce an integrated knowledge graph[1] such as the one depicted in Fig. 1.[2] Solving this problem requires semantic parsing of multiple natural language texts, transforming them to a formal representation, and identifying common vs. different parts in order to reason over an integrated knowledge graph associated with its textual provenance. Regardless of the chosen order, all these tasks must be addressed. Transforming natural language to formal representations has been investigated in ontology learning [3] and machine reading [4,5]; recognizing common parts in multiple sources is differently addressed by means of text similarity [6], co-reference resolution [7–9], ontology matching [10], and knowledge base integration [11,12].

In this paper, we describe in detail and experiment MERGILO, an improved version of the method proposed in [13] to handle multiple NL inputs, typically short text inputs such as news, in order to output knowledge graphs representing the integrated knowledge that they express. Integrating knowledge from multiple NL sources is crucial in order to implement intelligent applications requiring the ability to evolve a multi-source and dynamic knowledge base. However, this problem is challenging, considering that natural language can use heterogeneous forms for expressing similar knowledge. To complicate the situation, evaluation is also hard since no gold standards are available and not even universal standards for knowledge representations exist (different representations would require different gold standards). In this paper, after formally introducing the problem and presenting MERGILO, we describe how we built a gold standard (we will also refer to it as the ground truth), through a semi-automatic process, and

* Corresponding author.
E-mail address: misael.mongiovi@istc.cnr.it (M. Mongiovì).

[1] This paper refers to RDF/OWL[2] as the primary knowledge representation language for knowledge graphs.

[2] The picture is the result of a manual analysis and aims to show an ideal result that an automatic system should be able to approximate.

http://dx.doi.org/10.1016/j.knosys.2016.05.014
0950-7051/

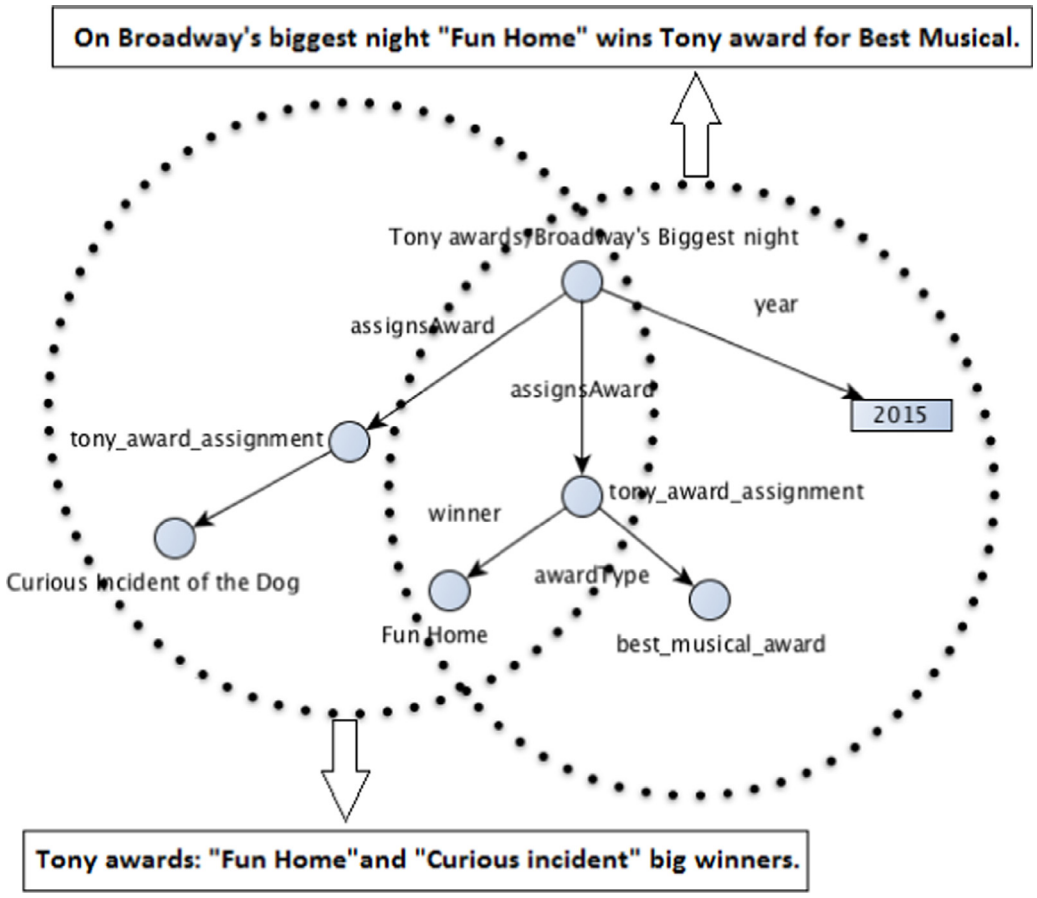

Fig. 1. An integrated formal representation resulting from knowledge acquired from two different news about 2015 Tony Awards. The link between text segments and extracted knowledge needs to be preserved.

starting from an existing annotated corpus for Cross-Document Coreference Resolution. The methodology is generalizable to other formal representations, and consists in generating a set of yes/no questions that can be answered by non-skilled people, using a crowd sourcing platform (in our case CrowdFlower[3]) to get the answers, and automatically generating the gold standard from the original corpus and the answers. In addition, we have tested our method against the generated gold standard, and compared the results to those produced by the existing baseline method.

Fig. 2 shows the overall pipeline of MERGILO: the input text sources are parsed and transformed into RDF knowledge graphs, then the knowledge graphs are *reconciled* by identifying their common parts. The first step is performed by reusing a state-of-the-art approach [5] (discussed in Section 3), while the second step is performed by means of a *knowledge reconciliation method* based on frame semantics and network alignment.

The rest of the paper is organized as follows: Section 2 discusses relevant related research. Section 3 presents our knowledge representation approach. Section 4 introduces our method to solve the knowledge reconciliation problem. Section 5 is dedicated to the evaluation of the proposed approach. Finally, Section 6 draws conclusions and shows directions where we are heading.

2. Related work

Cross-document Coreference resolution. The closest task to knowledge reconciliation, as defined in literature, is the NLP task known as Cross-document Coreference Resolution (CCR) [7]. CCR aims at associating mentions about a same entity (object, person, concept, etc.) across different texts. Relevant work addressing cross-document coreference resolution includes [14–16]. [7] uses spectral clustering and graph partitioning, and [17] is based on bag of words, latent similarity and clustering techniques. This problem is defined and solved in terms of text fragments, rather than formal constructs such as those composing a knowledge graph. Therefore the results of CCR are "extractive", and not applicable in "abstractive" tasks[4] that require a machine-usable representation of knowledge. Trying to transfer the knowledge from a CCR output to an abstract representation is hard. The identification of text fragments for annotating mentions is not unambiguously defined. For example in the sentence "People said Reid's representative Jack Ketsoyan confirmed..." of the EECB gold standard for CCR, the whole text fragment "Reid's representative Jack Ketsoyan" is considered a mention (which clearly refers to "Jack Ketsoyan"). However, parts of this text – taken alone – refer either to the same entity (e.g., "Jack", "representative", "Ketsoyan") or to other ones (e.g., "Reid"). Connecting that mention to the correct entity of an abstract representation is a non-trivial task that requires itself some degree of comprehension of the text. In contrast, solving the problem at an abstract level does not require handling text fragments, and has the further advantage of enabling the exploitation of additional information, including relations and semantic annotations, in order to improve the results.

[3] http://www.crowdflower.com/

[4] *Abstractive* means that the result of text analysis is not a (set of) text segment(s), but rather a *representation* of a text in a knowledge representation language, cf. [18] for a definition of *abstractive* techniques in NLP.