Scalable Understanding of Multilingual MediA (SUMMA)

http://www.summa-project.eu

H2020 Research and Innovation Action
Number: 688139
D4.2 – Release of multilingual named entity recognition, linking and coreference capabilities

Nature	Other	**Work Package**	WP4
Due Date	31/07/2017	**Submission Date**	31/07/2017
Main authors	Sebastião Miranda (PRIB)		
Co-authors	David Nogueira (PRIB), Afonso Mendes (PRIB)		
Reviewers	Guntis Barzdins (LETA)		
Keywords	Entity recognition, entity linking, coreference resolution		
Version Control			
v0.1	**Status**	Draft	29/06/2017
v0.2	**Status**	Draft	25/07/2017
v1.0	**Status**	Final	28/07/2017

Contents

Abstract

Named entity recognition, entity linking and coreference resolution are three Natural Language Processing (NLP) components developed in T4.1 (WP4) which process text in several languages and enrich it to enable automatic text understanding. In this deliverable we report the progress in these three components from a software tool release perspective and how these tools integrate in the SUMMA platform.

1 Introduction

Named entity recognition, entity linking and coreference resolution are three NLP components developed in T4.1 (WP4) which process and enrich text both from original media sources and also from the output of the Automatic Speech Recognition (ASR) and Machine Translation (MT) modules from WP3. The main goal of the current report is to describe these three components from a software tool release perspective, whereas a full research and development overview is to be reported in Deliverable D4.1. These component tools offer the following capabilities to the SUMMA platform:

- Automatic recognition of named entity mentions in the text (such as people, organisations, places and other types) in multiple languages.
- Linking of the recognised mentions of entities to a crosslingual knowledge base entity id.
- Solving coreference to help clustering (grouping) mentions of the same entity in a document, either by their name or via other referent such as a pronoun.

The detection and linking of named entities has a direct output to the platform interface, and helps a user not only to swiftly understand which entities are present in a given text but also to search the documents in the platform by a particular entity. Moreover, the output of these components may be used as a feature on several other SUMMA tasks, such as clustering (WP3), knowledge base population (WP4) and storyline summarization (WP5).

These components are currently deployed on standalone servers and integrate with the rest of the platform via a RESTful API, and will be subsequently shared as docker components (www.docker.com) within the SUMMA project to facilitate integration with the rest of the platform. A complete deployment status of these components is presented further in this report in Table 1.

Note: This document is not intended as a full technical description of the technologies developed in WP4 – that information is contained in Deliverable D4.1 (M18). The current deliverable details only the aforementioned first three components of WP4, namely the named entity recognition, entity linking and coreference resolution technologies developed in T4.1. This deliverable is related to Deliverable D8.3, which lists components for IP purposes at a coarser granularity.

2 Software release by component

2.1 Named Entity Recognition

Named entity recognition (or NER) is a popular research topic in the literature and consists of detecting entity mentions in a text (typically accompanied with an entity type). Within SUMMA, research efforts are actively being pursued to improve the current multilingual NER systems. Nevertheless, there are already production-ready systems currently integrated within the SUMMA project.

The current named entity recognition module is deployed as a RESTfull API for use within the SUMMA at http://213.63.185.148/EntityTagging/v2/. The SUMMA team carefully designed this flexible API interface and several object models in JSON schema in order to enable fast and scalable development across all SUMMA modules. This API receives a document (or a collection of documents) and tags detected entity mentions in the text. For Spanish and Portuguese the main underlying NLP module is Priberam's rule-based Text Analyser (Amaral et al. (2008)), whilst optionally also supporting English. Additionally, for English, German and Spanish the system implements a system based on Turbo Parser (Martins et al. (2013), Martins and Almeida (2014)), which uses a Conditional Random Field (CRF) sequence model whose features are based on the Illinois Entity Tagger (Ratinov and Roth, 2009). Currently this module performs at 11520 tokens/sec on a single core CPU. Both these NLP modules are integrated in a standalone server with the following characteristics:

- Standalone C# RESTful API server, with C++ production-ready core modules.
- Support for English, Spanish, Portuguese and German.

The released modules correspond to components 1-6 in Table 1. Since the named entity types depend on the available training data per language, we currently emit different types depending on the language. Currently our modules emit features of types such as people, organisations, places and others, whereas events are not addressed.

Emitting uniform entity types will be addressed in the future by using the entity types from the Entity Linking module (Wikipedia/FreeBase).

2.2 Entity Linking

The task of Entity Linking aims at connecting detected mention occurrences in one or multiple documents with known entity IDs in a knowledge base (e.g., Wikipedia, FreeBase). The main challenge resides in the fact that entities are not necessarily mentioned with their distinct and most complete knowledge base identifier (entity name). Therefore, statistics obtained from large document sets with links to knowledge base identifiers, as well as their contexts needs to be leveraged for an accurate disambiguation. Currently, we've been linking entities to their respective Wikipedia page ID (which is also aligned with a FreeBase ID) by outputting both a Wikipedia page link and a Freebase ID. In the context of the SUMMA project, two Entity Linking systems are currently being developed:

- The first is a production-ready Entity Linking system based on a nearest-neighbours ruled-based approach, using Priberam Search engine (Amaral et al., 2008). This system was submitted to the TAC KBP 2016 shared task (Paikens et al., 2017), and will be submitted again in 2017, with focus on the English and Spanish languages.

- Another system, which is still under research and development is an easy-first structured prediction approach to Entity Linking (see deliverable D4.1 for more detail). This second system aims to solve the Entity Linking task in a language-agnostic manner by leveraging multilingual representations with neural networks, and therefore might scale better to the remaining SUMMA languages.

The first system is currently available within SUMMA in the same API as the NER system (http://213.63.185.148/EntityTagging/v2/) in a standalone server with the following characteristics:

- Standalone C# RESTful API server, with C++ production-ready core modules.
- Support for English, Spanish, Portuguese and German.
- Performance (just entity linking component): 0.5 docs/sec on a multi-core CPU with 16GB RAM an SSD disk.

The released modules correspond to components 7-10 in Table 1.

2.3 Coreference Resolution

Coreference resolution is a NLP task which consists of clustering (grouping) mentions of the same entity in a document, either by their name or via other referent such as a pronoun. In the current release, we provide both an intra-document coreference system using Turbo Parser (Martins et al. (2013), Martins and Almeida (2014)) and an intra/inter-document system integrated in the Entity Linking module. These systems help the entity linking component disambiguate detected mentions in the text and also serve as features to other SUMMA modules (such as knowledge base construction in WP4). This module is currently deployed as a docker component for English and Spanish but it's not meant to be used as a standalone service. Instead, other NLP modules (such as Entity Linking) use this service internally. The released modules correspond to components 11 and 12 in Table 1.

3 Release summary

To better track down the current status of development, we define the following component categories:

- **Initial:** A working system, but not necessarily expected to work well in SUMMA due to small or mismatched training data.
- **Advanced:** A system that works well, and may also be expected to work well in SUMMA, for instance because it was trained on broadcast data.
- **Tuned:** A system that has been trained on or adapted to SUMMA-specific data.

Our current system implementations are either initial or advanced, currently covering the English, German, Spanish and Portuguese languages. The table below lists the tools that have been made available to the other partners either to support the development of other SUMMA technologies or to serve as direct outputs to the platform.

Name	Copyright	Licence	Docker
(1) English NER (TurboParser) Advanced	André Martins & Priberam URL: https://github.com/andre-martins/TurboParser	LGPL v3.0	Y
(2) Spanish NER (TurboParser) Advanced	André Martins & Priberam URL: https://github.com/andre-martins/TurboParser	LGPL v3.0	Y
(3) German NER (TurboParser) Initial	André Martins & Priberam URL: https://github.com/andre-martins/TurboParser	LGPL v3.0	Y
(4) English NER (PBA Analyser[1]) Initial	Priberam	SUMMA only	N
(5) Spanish NER (PBA Analyser[1]) Advanced	Priberam	SUMMA only	N
(6) Portuguese NER (PBA Analyser[1]) Advanced	Priberam	SUMMA only	N
(7) English Entity Linking Advanced	Priberam	SUMMA only	N
(8) Spanish Entity Linking Advanced	Priberam	SUMMA only	N
(9) German Entity Linking Advanced	Priberam	SUMMA only	N
(10) Portuguese Entity Linking Advanced	Priberam	SUMMA only	N
(11) English Coreference Resolution Advanced	André Martins & Priberam URL: https://github.com/andre-martins/TurboParser	LGPL v3.0	Y
(12) Spanish Coreference Resolution Advanced	André Martins & Priberam URL: https://github.com/andre-martins/TurboParser	LGPL v3.0	Y

Table 1: Named Entity Recognition and Linking (NER,NEL) and coreference components.

Since we are still in a development phase, more than one different system is available for some tasks. Further in the development of the project, we plan to integrate at least one named entity recognition and one entity linking system in a docker component for English, Spanish, German and Portuguese.

[1] Priberam's rule-based Text Analyzer (Amaral et al., 2008)

4 Future plans

The current software release already offers a reasonable coverage of features to enable development of the SUMMA platform towards it's ultimate goals. In order to approach this objective, we plan to continue improving our current systems in future software releases. To this end, we envisage five classes of future actions:

- **First priority: Moving more components into the Docker architecture.** Some components are not yet available as Docker modules; this is indicated in the component table above. Release of a component as a Docker module is taken as that component being usable by the wider project. We plan to have these components released as docker modules in 2017:
  - Entity Linking: English, German, Spanish and Portuguese
  - Named Entity Recognition: English, German, Spanish and Portuguese
- **Second priority: Iterative development of better models.** We'll continue to pursue the research and development of better models for named entity recognition and entity linking, in particular the neural approaches described in this document and further detailed in deliverable D4.1.
- **Third priority: Coordination with ASR/MT streaming modules** The current and planned NER implementation modules expect properly capitalized text as the input. Presently, the ASR modules in the SUMMA platform output non-capitalised text, which is not compatible with the follow-up NER processing. Proper capitalization and punctuation in the ASR output could be inserted by the separate Punctuation module or by the MT module in order to mitigate this problem. Another option would be to make an addition to the ASR modules by devising a scheme to feedback the entities recognized by the NER modules in proper news text (written news articles) into the ASR modules in order to help recognize new entities present in the audio streams.
- **Fourth priority: Improvement of computational performance** We'll continue to improve the computational performance of these components, in particular the Entity Linking modules.

References

Carlos Amaral, Adán Cassan, Helena Figueira, André Martins, Afonso Mendes, Pedro Mendes, José Pina, and Cláudia Pinto. Priberam's question answering system in qa@ clef 2008. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 337–344. Springer, 2008.

A. F. T. Martins, M. B. Almeida, and N. A. Smith. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, 2013.

André F. T. Martins and Mariana S. C. Almeida. Priberam: A turbo semantic parser with second order features. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 471–476. Association for Computational Linguistics, 2014.

Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning - CoNLL*, pages 147–155, 2009.

Peteris Paikens, Guntis Barzdins, Afonso Mendes, Daniel Ferreira, Samuel Broscheit, Mariana S. C. Almeida, Sebastião Miranda, David Nogueira, Pedro Balage, and André F. T. Martins. Summa at tac knowledge base population task 2016. In *Proceedings of the Text Analysis Conference -TAC*, pages 1–9, Gaithersburg, Maryland USA, 2017.

ENDPAGE

SUMMA

H2020-ICT-2015 688139

D4.2 Release of multilingual named entity recognition, linking and coreference capabilities