Neji: a BioC compatible framework for biomedical concept recognition

Sérgio Matos, André Santos, David Campos, and José Luís Oliveira

IEETA/DETI, University of Aveiro,
Campus Universitrio de Santiago, 3810-193 Aveiro, Portugal
{aleixomatos,andre.jeronimo,jlo}@ua.pt
BMD Software, 3810-074 Aveiro, Portugal
david.campos@bmd-software.com

Abstract. The BioCreative V Collaborative Biocurator Assistant Task aimed at promoting the development of BioC-compatible modules that can be integrated into a text-mining assisted biocuration tool. This paper describes the Neji framework for biomedical concept recognition and its methods, based on machine-learning, dictionary-matching and post processing rules, used for our participation in the first three tasks: gene mention recognition, organism name recognition and normalization, and gene name normalization.

Key words: Biomedical concept recognition, Biocuration, BioC format

1 Introduction

This paper describes our participation in the BioCreative V Collaborative Biocurator Assistant Task (BioC). BioC is a simple format to share text data and annotations that allows a large number of different annotations to be represented [4]. Annotation semantics in BioC are user-defined and are represented through a human-readable key file that indicates, for example, the BioC infon used to represent the normalized identifier for an annotation. The purpose of this task is to create complementary BioC-compatible modules that can be integrated into a system to assist biocurators, namely for identification of protein-protein molecular interaction information for the BioGRID [3] database.

We participated on the following subtasks of the BioC challenge:

- Task 1: Gene/protein NER - identify gene/protein mentions in text
- Task 2: Species/organism NER - identify mentions of species/organism names and normalize to NCBI Taxonomy identifiers
- Task 3: Gene/protein normalization - assign Entrez Gene IDs to gene/protein mentions based on species/organisms mentioned in the text

2 Systems description and methods

We used Neji [1], a modular biomedical concept recognition framework that provides state-of-the-art methods for natural language and text processing (sentence

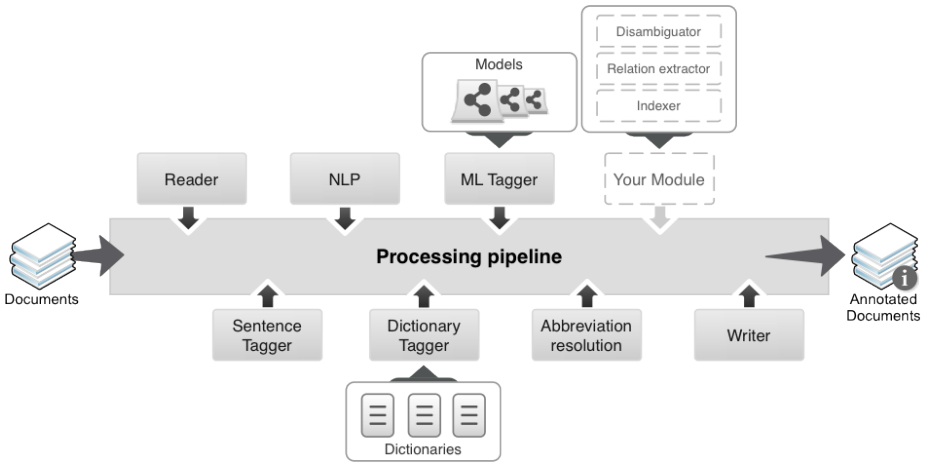

Fig. 1. Architecture of the Neji framework.

splitting, tokenization, lemmatization, part-of-speech tagging, chunking and dependency parsing), concept recognition through efficient dictionary-matching and machine learning models, and multi-threaded document processing. Figure 1 illustrates the modular architecture of Neji. Neji can be used as an off-the-shelf tool for named entity recognition (NER) or concept recognition (with normalization), either through the source code or by using the simple command-line interface (CLI). Additionally, new modules can be created following the framework's API and integrated in the processing pipeline. Neji supports various input and output formats, namely Pubmed and PubMed Central XML, IeXML, CoNLL, A1 and BioC, and others can be added by developing special purpose reader and/or writer modules.

Task 1: Gene/protein NER

For the Gene/protein NER task we applied a second-order conditional random fields (CRF) model that achieves a recognition performance of 87% in F-score when evaluated on the BioCreative II gene mention recognition corpus [7]. After recursive feature selection, the following optimized feature set was used to train the model, as described in [2]:

- Token and NLP features:
  - Token, lemma, POS, chunk tags.
- Orthographic features:
  - Capitalization (e.g., "InitCap" and "AllCaps");
  - Digits and capitalized characters counting (e.g., "TwoDigit" and "TwoCap");
  - Symbols (e.g., "Dash", "Dot" and "Comma");
  - Greek letters (e.g., features for "alpha" and "α").
- Morphological features:

    - Prefixes, suffixes, and 2, 3 and 4 character n-grams;
    - Word shape features to reflect how letters, digits and symbols are organized in the token (e.g., the structure of "Abc:1234" is expressed as "Aaa#1111").
- Domain knowledge:
    - Dictionary matching using the BioLexicon dictionary.
- Local context:
    - Tokens, lemmas and dictionary matching features in the window -3,3, plus all the features in the window -1,1.

Task 2: Species/organism NER

For the Species/organism NER task we used the dictionary provided by LINNAEUS [5]. To eliminate ambiguous cases, namely when the short species name was used, we first identified all unambiguous organism annotations across the full text. Then, for each ambiguous organism mention, we check if one of the possible identifiers was unambiguously assigned to another mention in the text, and select that identifier. If the ambiguity was not solved in this step, we then check against the synonyms of the most common model organisms[1], and select the matching identifier.

Task 3: Gene/protein normalization

For gene name normalization, we used a dictionary lookup strategy in which each dictionary from a prioritized list is checked in sequence, as described in [1]. We used two dictionaries created from the BioLexicon gene dictionary [6], the first one containing only the preferred name of each gene, and the second one containing all the synonyms. Following our strategy, the first dictionary is searched for a matching identifier for the gene mention. If a match is found the algorithm finishes, otherwise the second dictionary is searched.

To filter out ambiguous cases, that is, gene mentions for which two or more entries were found in the dictionaries, we applied a series of rules based on unambiguous gene annotations and species annotations in the text. We start by identifying all unambiguous gene annotations across the full text, and then, for each ambiguous gene mention we apply the following sequence of rules:

1. Unambigous synonym in text:
    - If one of the possible identifiers for a gene mention was unambiguously assigned to another mention in the text, that identifier is selected.
2. Species in sentence context:
    - If a species annotation exists in the same sentence, we select from the list of possible gene identifiers, the ones that belong to the mentioned species[2];

[1] As listed in: `http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi`

- If only one gene identifier is left, we add this to the list of unambiguous gene annotations, to be used in the next steps.

3. Species in passage context:
   - If a species annotation exists in the same passage, we select from the list of possible gene identifiers, the ones that belong to the mentioned species[2];
   - If only one gene identifier is left, we add this to the list of unambiguous gene annotations, to be used in the next steps.
4. Species in document context:
   - If a species annotation exists in the same document, we select from the list of possible gene identifiers, the ones that belong to the mentioned species[2].

3 Conclusions

We presented our concept recognition methods for the first three subtasks of the BioCreative V Biocurator Assistant Task, namely gene/protein NER, organism NER, and gene name normalization. Using Neji, a modular framework for biomedical concept recognition, we applied a machine-learning approach based on CRFs for the recognition of gene mentions, and dictionary-matching methods for the identification of organism names, with post-processing rules to reduce ambiguity and normalize gene mentions.

Acknowledgments. This work was supported by national funds through FCT - Foundation for Science and Technology, in the context of the projects UID/CEC/00127/2013 and Incentivo/EEI/UI0127/2014. D. Campos has received support from the HemoSpec European project (EC contract number 611682). S. Matos is funded by FCT under the FCT Investigator programme.

References

1. Campos, D., Matos, S., Oliveira, J.L.: A modular framework for biomedical concept recognition. BMC bioinformatics 14(281) (2013)
2. Campos, D., Matos, S., Oliveira, J.L.: Gimli: open source and high-performance biomedical name recognition. BMC bioinformatics 14(1), 54 (2013)
3. Chatr-Aryamontri, A., Breitkreutz, B.J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., ODonnell, L., et al.: The biogrid interaction database: 2013 update. Nucleic acids research 41(D1), D816–D823 (2013)
4. Comeau, D.C., Doğan, R.I., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., et al.: Bioc: a minimalist approach to interoperability for biomedical text processing. Database 2013, bat064 (2013)
5. Gerner, M., Nenadic, G., Bergman, C.M.: LINNAEUS: a species name identification system for biomedical literature. BMC bioinformatics 11, 85 (2010)

[2] If no species mention was found anywhere on the full document, we select the gene identifier(s) that belong to the human species.

6. Sasaki, Y., Montemagni, S., Pezik, P., Rebholz-Schuhmann, D., McNaught, J., Ananiadou, S.: Biolexicon: A lexical resource for the biology domain. In: Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008) (2008)
7. Smith, L., Tanabe, L.K., Ando, R.J., Kuo, C.J., Chung, I.F., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C.M., Ganchev, K., et al.: Overview of biocreative ii gene mention recognition. Genome biology 9(Suppl 2), S2 (2008)