Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report accompanying Manually Annotated Reference Corpus

Deliverable number	*D1.4.1*
Dissemination level	*Public*
Delivery date	*31st May 2013*
Status	*Final*
Author(s)	*Angus Roberts, Johann Petrak, Niraj Aswani*

This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Abstract

The Khresmoi project is building a multi-lingual search and access system for biomedical information and documents. The project is using automatic recognition of medical entities in text, such as Disease and Drugs, to assist with that search. Automatic recognition of these entities is trained by manual correction of machine annotations. This correction leads to the construction of a corpus of manually corrected documents, which may be of use in other biomedical NLP tasks.

Table of Contents

1 Executive summary

The Khresmoi project is using automatic recognition of medical entities in text, such as Disease and Drugs, to assist with search of biomedical documents. Automatic recognition is trained iteratively using manually correction of automatic annotations in text. A side effect of this process is the production of a corpus of manually corrected documents, that may be of use in other biomedical language processing tasks.

The corpus consists of two parts. The first part is an initial corpus constructed during the early part of the Khresmoi project, the second part is the ongoing collection of annotations as part of the iterative training of the entity recognition system. This report describes these two parts: their structure; the document format used; the annotations contained within documents; and gives statistics describing the corpus.

Khresmoi is improving the performance of entity recognition by coupling automatic recognition with manual correction, in an iterative machine learning approach. The manual data being collected as a result of that correction, and which accompanies this report, is a valuable resource beyond Khresmoi, of potential use in training and evaluating other biomedical language processing applications.

2 Introduction

This report accompanies Khresmoi deliverable D1.4.1, Manually Annotated Reference Corpus, and describes the content of that deliverable. The reference corpus is a collection of Khresmoi web documents with key entities (such as disease, drug) marked (annotated), in the text. It has been constructed by first annotating these entities with an imperfect automatic process. These automatic annotations have then been corrected by several human operators, and the differences between these human operators resolved to give the final reference corpus.

2.1 Background

The Khresmoi project is building a multi-lingual search and access system for biomedical information and documents [14]. Several technologies are used to improve search, including machine recognition of medical entities within text, and linking of these to the Khresmoi Knowledge Base. Machine recognition of entities requires training data, and this is provided by the Manually Annotated Reference Corpus accompanying this report.

Entity recognition and corpus creation are coupled iteratively, in that:

1. an initial machine annotated corpus is created;
2. this is corrected by human annotators;
3. the machine entity recognition model is updated with these corrections
4. the steps are repeated.

This process and the results from running it are reported in the sibling report D1.4.2, “Report on Coupling Manual and Automatic Annotation” [6].

Manual annotations are corrected according to a set of guidelines, described in Khresmoi deliverable D1.1 “Manual Annotation Guidelines and Management Protocol”[5], which were themselves based on

the project requirements, described in Khresmoi Deliverable D8.2 "Use case definition including concrete data requirements" [7].

An early version of the system used to create automatic annotations was described in D1.2 "Initial prototype for semantic annotation of the Khresmoi literature" [2], and early results of the manual annotation described in D1.3 "Report on Results of the WP1 First Evaluation Phase" [1].

2.2 Summary of this report

The description of the reference corpus given in this report consists of two parts. The first, described below in the section "Initial corpus", was constructed while Khresmoi project requirements were elucidated. With the second part, described under "Ongoing corpus", construction was started after requirements were complete. Construction of this second corpus is ongoing. The delivered corpus, and this accompanying report, will be updated as further data becomes available.

3 Initial corpus

This section describes the initial corpus: what documents it is composed of; the document format; a description of the annotation schema (i.e. what annotations are present in documents, and what information is associated with them); and the corpus content, as high level document and annotation statistics.

3.1 Corpus description

The original corpus consisted of 883 documents randomly drawn from the Gene Home Reference [8]. These were automatically annotated by the application described in D1.2 "Initial prototype for semantic annotation of the Khresmoi literature" [2] , to mark anatomical locations and medical problems. Three annotators corrected each document independently, and a fourth annotator created a "consensus" set, checking and resolving differences between annotators.

For some of these 883 documents, one or more annotators failed to provide any Anatomy or Problem annotations. These were removed from the final corpus, to leave 625 documents.

3.2 Document format

Documents are provided in GATE XML format, a standard based stand off markup. This format is described in [9]. Documents may either be opened directly in a desktop tool, GATE Developer, or may be opened via the GATE Embedded Library API, and manipulated programmatically [10,11]. Documents may be converted to other formats either via the GATE Developer menus, or using GATE Embedded Library. New formats may be trivially supported by writing an exporter for the GATE Embedded Library [12].

3.3 Schema

Documents are structured in the following way:

- A document may contain several annotation sets: i.e. named sets grouping several annotations together, for example all annotations created by a single annotator
- Each annotation set may contain several named types of annotation, such as "Disease" or "Anatomy"
- Each annotation may contain a map of feature names to values

The annotations in the initial corpus are organised as follows:

- Annotation sets: all documents may contain these sets of annotations:
  - preprocessing: annotations added by automatic annotation, for correction
  - annotator1: corrections of the preprocessing set by an annotator
  - annotator2: corrections of the preprocessing set by a second annotator
  - annotator3: corrections of the preprocessing set by a third annotator
  - consensus: manual reconciliation of sets annotator1,2,3
- Annotations: all of the above annotation sets may contain these annotations:
  - Section: splitting the document in to sections that contain content, as opposed to sections that contain administrative, presentation and other such material (for example, disclaimers and web page menus).
  - Problem: an annotation describing a disease, medical problem, etc. These are only found inside Section annotations.
  - Anatomy; an annotation describing an anatomical location. These are only found inside Section annotations.
- Features: the Anatomy and Problem annotations are assigned a type from the Khresmoi Knowledge Base, using type identifiers from the UMLS semantic network [13]. The following types are used:
  - Anatomy: T023 (Body Part, Organ, or Organ Component), T029 (Body Location or Region), T030 (Body Space or Junction)
  - Problem: T047 (Disease or Syndrome), T048 (Mental or Behavioral Dysfunction), T191 (Neoplastic Process), T020 (Acquired Abnormality), T037 (Injury or Poisoning), T190 (Anatomical Abnormality)

3.4 Content

The corpus contains documents with the following characteristics:

Number of documents		625
Number of tokens in documents (excluding markup)	**Minimum**	26
	Median	727
	Mean	1039.8
	Maximum	8306

The number of semantic annotations in these documents is as follows:

Annotation	**Semantic type**	**Semantic type label**	**Number of annotations with this type**
Problem	T047	Disease or Syndrome	9973
	T048	Mental or Behavioral Dysfunction	235
	T191	Neoplastic Process	986
	T020	Acquired Abnormality	67
	T037	Injury or Poisoning	66
	T190	Anatomical Abnormality	83
Anatomy	T023	Body Part, Organ, or Organ Component	1147
	T029	Body Location or Region	409
	T030	Body Space or Junction	78

4 Ongoing corpus construction

This section describes the ongoing corpus, i.e. that portion of the corpus that is still being collected. The section describes what documents it is composed of; the document format; a description of the annotation schema (i.e. what annotations are present in documents, and what information is associated with them); and the corpus content, as high level document and annotation statistics.

As this portion of the corpus is still being collected, the description is incomplete, in that further documents and annotation of those documents will be added as they become available. As the manually corrected corpus grows, further documents will be added.

4.1 Corpus description

The ongoing corpus currently consists of two sets of documents, equal numbers of which have been automatically annotated, and are being manually corrected:

- CouchDb documents: i.e. documents drawn from the full Khresmoi web crawl, which are stored in a CouchDb database by the Khresmoi Health On the Net partner (HON). In the corpus, the names of these documents are prefixed with the letter "W".
- Wikipedia medical pages. Wikipedia does not allow crawling, and so these are therefore downloaded separately. In the corpus, the names of these documents are prefixed with the letter "K".

In both cases, documents are selected to meet useful criteria for manual annotation: English documents of a length that can be done by a human annotator in 15 to 30 minutes, which experimentation with annotators has shown to be between 200 and 2000 non-markup tokens, i.e. tokens that are not HTML tags.

4.1.1 CouchDB document preparation

1. Retrieved 19000 documents from CouchDB by language "en", and with the highest relevancy as given by a score calculated by HON. This score is provided with each document record in the HON database.
2. Filtered document with between 200 and 2000 Tokens which leaves us with 6950 documents.
3. These documents are processed with the GATE application described in Deliverable D1.4.2 to give the a set of automatic annotations, as described in the "Schema" section below, and are currently being corrected by manual annotators.

4.1.2 Wikipedia document preparation

1. A Wikipedia dump of medically relevant English language pages, as used in Khresmoi, was converted to 10284 individual GATE document files where the document name was created in the form "K_" concatenated with the URL-encoded title of the page.
2. From this, 5518 documents with 200..2000 Tokens were selected.

3. These documents are processed with the GATE application described in Deliverable D1.4.2 to give the a set of automatic annotations, as described in the "Schema" section below, and are currently being corrected by manual annotators.

4.2 Document format

Documents are provided in GATE XML format, as described for the Initial Corpus.

4.3 Schema

In the same way as for Initial Corpus, documents are structured to contain annotations grouped into sets. The annotations in the ongoing corpus are organised as shown below. Note that in the ongoing exercise the use of a consensus annotator to resolve annotator differences, has been replaced with the use of larger numbers of individual annotators, the intention being to use a voting strategy for differences.

- Annotation sets: all documents may contain these sets of annotations:
  - preprocessing: annotations added by automatic annotation, for correction
  - annotator1-7: corrections of the preprocessing set by up to seven annotators
- Annotations: all of the above annotation sets may contain these annotations:
  - Content and Non-content: splitting the document in to sections that contain content, as described earlier.
  - Disease: an annotation describing a disease, medical problem etc. These are only found inside Content annotations.
  - Anatomy: an annotation describing an anatomical location. These are only found inside Content annotations.
  - Drug: an annotation describing a prescribed drug. These are only found inside Content annotations.
  - Investigation: an annotation describing a medical investigation. These are only found inside Content annotations.
- Features: the following features are added to pre-processed document. Annotations added by annotators will have no features.
  - comment: an optional comment, that may be added by annotators
  - cui: Khresmoi Knowledge Base instance identifier, based on the UMLS concept identifier for the marked entity
  - tui: Khresmoi Knowledge Base instance identifier, based on the UMLS semantic type identifier for the marked entity
  - tui-name: The human readable name of the UMLS semantic type for the marked entity

4.4 Content

The corpus contains documents with the following characteristics:

		CouchDb	Wikipedia
Number of documents		6950	5518
Characters in original HTML:	**Minimum**	2148	1207
	Median	74387	6436
	Mean	81312	7794
	Maximum	634961	62722
Non-markup characters, as parsed out by GATE	**Minimum**	891	773
	Median	10118	3556
	Mean	12059	4302
	Maximum	92568	30428
Total tokens:	**Minimum**	215	200
	Median	1918	651
	Mean	2410	789
	Maximum	17636	2000
Tokens within content annotations	**Minimum**	200	200
	Median	768	651
	Mean	836	789

	Maximum	2000	2000

4.5 Current status of manual annotation

The manual correction of the above documents is ongoing. Sixty annotators have been recruited and trained, of which 14 are actively annotating so far. Each document in the corpus is being corrected by five annotators. The table below shows the documents that have been distributed to annotators so far, and how many annotators have completed these documents. The total documents column gives the total number of sets of document corrections. So for example, if a document has so far been corrected 3 times, it will contribute 3 to this column.

Number of documents	**Number of annotators correcting these document so far:**	**Total documents**
20	5	100
5	4	20
6	3	18
41	2	82
6	1	6
Total number of documents annotated:		**226**

5 Conclusion

In order to build a multi-lingual search and access system for biomedical information and documents, the Khresmoi project requires machine-based recognition of medical entities. Machine recognition of entities is being achieved by coupling automatic recognition of entities with manual correction, in an iterative machine learning approach. This report has described the manual data being collected as a result of that correction, and which accompanies this report. The manual data is a valuable resource beyond Khresmoi, of potential use in training and evaluating other biomedical language processing applications.

6 References

[1] Niraj Aswani, Liadh Kelly, Mark Greenwood, Angus Roberts, Matthias Samwald, Natalia Pletneva, Gareth Jones, Lorraine Goeuriot. Report on Results of the WP1 First Evaluation Phase, Khresmoi project deliverable D1.3, August 2012.

[2] Mark A. Greenwood, Angus Roberts, Niraj Aswani, Phil Gooch. Initial prototype for semantic annotation of the Khresmoi literature, Khresmoi project deliverable D1.2, May 2012.

[3] Betsy L. Humphreys and Donald A.B. Lindberg and Harold M. Schoolman and G. Octo Barnett. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 1998, 5:1

[4] King, B., Wang, L., et al. (2011). Cenagage Learning at TREC 2011 Medical Track. The Twentieth Text Retrieval Conference Proceedings (TREC 2011), Gaithersburg, MD. National Institute for Standards and Technology.

[5] Angus Roberts, Niraj Aswani, Natalia Pletneva, Celia Boyer, Thomas Heitz, Kalina Bontcheva, Mark A. Greenwood. Manual Annotation Guidelines and Management Protocol, Khresmoi project deliverable D1.1, February 2012.

[6] Angus Roberts, Johann Petrak, Niraj Aswani, Report on Coupling Manual and Automatic Annotation, Khresmoi project deliverable D1.4.2, May 2013

[7] Use case definition including concrete data requirements. Khresmoi project deliverable D8.2

[8] Genetics Home Reference: http://ghr.nlm.nih.gov/

[9] S. Bird and M. Liberman. A Formal Framework for Linguistic Annotation. Technical Report MS-CIS-99-01, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, 1999. http://xxx.lanl.gov/abs/cs.CL/9903003.

[10] H. Cunningham, et al. Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. 15 April 2011. ISBN 0956599311

[11] H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva (2013) Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol 9(2): e1002854. doi:10.1371/journal.pcbi.1002854

[12] GATE Javadocs, available from http://gate.ac.uk

[13] Semantic Network: http://www.ncbi.nlm.nih.gov/books/NBK9679/

[14] A. Hanbury, C. Boyer, M. Gschwandtner, H. Müller. KHRESMOI: towards a multi-lingual search and access system for biomedical information. Med-e-Tel, Luxembourg, 2011.