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	*Draft*
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

Index of Tables

List of abbreviations

API	Application Programming Interface
CUI	Concept Unique Identifier
DB	Database
F1	Harmonic mean of precision and recall, weighted equally for both precision and recall
GATE	General Architecture for Text Engineering
HON	Health on the Net
HTML	Hyper Text Markup Language
IAA	Inter Annotator Agreement
ML	Machine Learning
NLP	Natural Language Processing
P	Precision
R	Recall
RadLex	Radiology Lexicon
TUI	Term Unique Identifier (UML ID)
UMLS	Unified Medical Language System
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
USFD	University of Sheffield
UTS	UMLS Terminology Service

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 collection of annotations created 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 “Iterative training corpus”, construction was started after requirements were complete.

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 1083 documents (previously reported as 883 document, in error) 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 the documents, one or more annotators failed to provide any Anatomy or Problem annotations. For others, no Section annotation was present. Documents with these errors were removed from the final corpus, to leave 625 documents.

3.2 Document format

Documents are provided in GATE XML format, a standards 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:
  - 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 ConceptId feature from the Khresmoi Knowledge Base, using concept identifiers (CUIs) from the UMLS semantic network [13]. These were restricted to the following UMLS types:
  - 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

Table 1: Initial corpus, statistics

The number of semantic annotations in these documents can be derived by mapping the annotation ConceptId feature to the UMLS Semantic type, giving the following frequencies:

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

Table 2: Initial corpus, annotation counts

4 Iterative training corpus

This section describes the iterative training corpus i.e. that portion of the corpus used for development of the Khresmoi annotation application. 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.

4.1 Corpus description

The corpus was constructed by first taking a set of source documents from the Khresmoi document repository and the web. We will refer to this sampled set, not annotated, as the source corpus. This was further sampled to give sub-sets for automatic annotation and manual correction. We will refer to these sub-sets as the annotated corpora.

4.1.1 Source corpora

The source corpus was drawn from two sets of documents. Only some of these have been automatically annotated and 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 were 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.

4.1.1.1 CouchDB document preparation

CouchDb documents were prepared as follows:

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 left us with 6950 documents which formed our reservoir of Web documents.

4.1.1.2 Wikipedia document preparation

Wikipedia documents were prepared as follows:

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 which formed our reservoir of Wikipedia documents.

4.1.2 Annotated corpora

For each manual annotation iteration, a corpus was prepared by randomly sampling 100 Wikipedia and documents and 100 other web documents from the set of selected documents. Sampled documents were removed from the reservoir so they would not be included in more than one corpus. Six corpora were prepared in this way.

The documents in the corpus were then 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. These annotations were then corrected by annotators. All documents with corrections by 5 annotators were selected, and a consensus annotation set created from the all annotations where there was majority agreement on the correction, for co-extensive annotations, using the GATE AnnotationMerging processing resource.

Annotation made use of UMLS as downloaded from the Khresmoi Knowledge Base on March 18th 2013.

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:
  - annotator1-5: 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

4.4.1 Source corpora

The source 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	**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 in content annotations	**Minimum**	200	200
	Median	768	651
	Mean	836	789
	Maximum	2000	2000

Table 3: Source corpus, statistics

4.4.2 Annotated corpora

The annotated corpora are described in the following three tables:

- Numbers of annotations in the consensus annotation set
- Numbers of annotations created by individual annotators
- Inter annotator agreement for each corpus

4.4.2.1 Numbers of annotations

Corpus	Documents	Tokens	Anatomy	Disease	Drug	Investi-gation
C0504	192	296178	1832	4861	2844	1408
C0512	160	188195	1734	5032	2792	1243
C0524	126	147703	1411	4173	1943	675
C0701	198	292150	1560	4349	1978	475
C0830	139	159396	1046	3125	1560	265
C0925	100	80398	732	1962	1256	116
Total	915	1164020	8315	23502	12373	4182

Table 4: Annotated corpora

Corpus	Annotator	Documents	Tokens	Content Tokens	Anatomy	Disease	Drug	Investigation
C0504	annotator1	192	296178	151383	2240	5552	3153	1462
	annotator2	192	296178	152858	2214	5564	3222	1758
	annotator3	192	296178	152167	1997	5448	3024	1568
	annotator4	192	296178	155419	1833	5434	2749	1262
	annotator5	192	296178	155585	1900	5320	2894	1319
C0512	annotator1	160	188195	122084	1774	4999	2010	483
	annotator2	160	188195	121399	1871	5881	2480	1053
	annotator3	160	188195	118849	1697	5314	2744	1150
	annotator4	160	188195	120834	1676	4748	2413	1037
	annotator5	160	188195	119931	1765	5116	2800	1078
C0524	annotator1	126	147703	103157	1383	4101	1803	520
	annotator2	126	147703	103998	1399	4360	1876	559
	annotator3	126	147703	103136	1478	4563	2005	684
	annotator4	126	147703	103321	1410	4352	1820	597
	annotator5	126	147703	103512	1425	4400	2038	616
C0701	annotator1	198	292150	150406	1654	4833	2140	510
	annotator2	198	292150	151208	1548	4624	1973	360
	annotator3	198	292150	149589	1568	4520	1667	407
	annotator4	198	292150	149360	1633	4727	2025	572
	annotator5	198	292150	149140	1592	4555	1819	465
C0830	annotator1	139	159396	108704	1167	3386	1542	181
	annotator2	139	159396	108680	1175	3644	1763	234
	annotator3	139	159396	108121	1184	3424	1562	238
	annotator4	139	159396	108549	1019	3108	1376	149
	annotator5	139	159396	108403	1088	3181	1403	149
C0925	annotator1	100	80398	80398	613	1771	907	65
	annotator2	100	80398	80398	757	2107	930	68
	annotator3	100	80398	80398	886	2331	1217	91
	annotator4	100	80398	80398	646	1750	916	57
	annotator5	100	80398	80398	837	2047	949	83

Table 5: Annotated corpora analysed by individual annotators

4.4.2.2 Inter-Annotator Agreement

Inter-Annotator Agreement (IAA) is the overall F measure between all pairings of annotators. For example, if there were three annotators A, B and C, then the three agreements between A and B, A and C, B and C would be averaged.

Two averages are given: micro and macro average:

Micro average: The micro averages were calculated over all annotator pairs, annotation types and documents, separately for strict and lenient matches.

Macro average: First the micro averages for each type over all annotator pairs and over all documents were calculated. The macro F measures were then calculated as the arithmetic mean of the micro averaged F measures for all annotation types. This was again done separately for strict and lenient matches.

Corpus	**Number of documents**			**Micro averaged IAA**		**Macro averaged IAA**	
	Total	**Wikipedia**	**Rest of web**	**Strict**	**Lenient**	**Strict**	**Lenient**
C0504	192	94	98	0.62	0.68	0.61	0.66
C0512	160	98	62	0.69	0.72	0.65	0.68
C0524	126	98	28	0.76	0.79	0.72	0.74
C0701	198	100	98	0.73	0.77	0.67	0.71
C0830	139	99	40	0.75	0.78	0.68	0.71
C0925	100	100	0	0.66	0.70	0.58	0.62

Table 6: Annotated corpora: inter annotator agreement

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 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 infromation. Med-e-Tel, Luxembourg, 2011.