Annotating Cohesive Statements of Anatomical Knowledge Toward Semi-automated Information Extraction

Kazuo Hara[1], Ikumi Suzuki[1], Kousaku Okubo[1] and Isamu Muto[2]
[1]*National Institute of Genetics, Mishima, Shizuoka, Japan*
[2]*BITS. Co., Ltd., Chiyoda, Tokyo, Japan*

Keywords: Semi-Automated Information Extraction, Cohesive Text, Itemized Text.

Abstract: Anatomical knowledge written in a textbook is almost completely unreusable computationally, because it is embedded in a cohesive discourse. In discourse contexts, the frequent use of cohesive ties such as reference expressions and coordinated phrases not only troubles the function of automated systems (i.e., natural language parsers) to extract knowledge from the resulting complicated sentences, but also affects the identification of mentions of anatomical named entities (NEs). We propose to revamp the prose style of anatomical textbooks by transforming cohesive discourse into itemized text, which can be accomplished by annotating reference expressions and coordinating conjunctions. Then, automatically, each anaphor will be replaced by its antecedent in each reference expression, and the conjoined elements are distributed to sentences duplicated for each coordinating conjunction connecting phrases. We demonstrate that, compared to the original text, the transformed one is easy for machines to process and hence convenient as a way of identifying mentions of anatomical NEs and their relations. Since the transformed text is human readable as well, we believe our approach provides a promising new model for language resources accessible by both human and machine, improving the computational reusability of textbooks.

1 INTRODUCTION

Scientific textbooks are language resources in which the knowledge continually being accumulated by humankind is presented to readers in order to elucidate the nature of the universe. Compared to genres of text such as news or weblogs, which have commonly been exploited by natural language processing researchers, the density of valuable knowledge in textbooks is actually considerably higher, because they are a vehicle for the overview of consensus knowledge carefully ascertained by the scientific community, packed tightly sentence by sentence. For instance, anatomical textbooks, which we will focus on here, describe the structures of the human body, the composition of individual tissues, and the relations between them. Such anatomical knowledge is fundamental for good health care, in that medical activities such as diagnosis, intervention, and prognosis aim to cure structures or tissues of the human body from ailments; hence, detailed knowledge of those structures is needed.

However, these concentrated accumulations of knowledge are at present almost totally unreusable computationally. In particular, they are not suited to automatic processing for semantic searching and reasoning. This is because the knowledge in a textbook is embedded in a cohesive discourse. In other words, a textbook does not comprise a collection of separate pieces of knowledge like subject–predicate–object triples in a resource description framework (RDF), which are employed in Linked Open Data (LOD) or knowledge bases such as FreeBase and DBPedia. Instead, what we do with a textbook is just read and understand it in our mind.

Thus, it seems that isolating individual pieces of knowledge from anatomical textbooks should be useful. However, this is not a straightforward process, because description is highly cohesive. As a result of the exigency to reduce redundancy using cohesive ties (that is, reference expressions, coordinated phrases), most statements in a given context depend on each other, and often do not make sense without reference to each other. In fact, it has been reported that cohesive ties are frequently used in scientific texts (Schäfer et al., 2012). This is likely because it is thought that the coherence yielded by the ties increases the quality of the text (Witte and Faigley, 1981).

In order to overcome difficulties arising from the

Hara K., Suzuki I., Okubo K. and Muto I..
Annotating Cohesive Statements of Anatomical Knowledge Toward Semi-automated Information Extraction.
DOI: 10.5220/0005132303420347
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2014), pages 342-347
ISBN: 978-989-758-048-2

The sclera has received its name from its extreme density and hardness; it is a firm, unyielding membrane, serving to maintain the form of the bulb. It is much thicker behind than in front; the thickness of its posterior part is 1 mm. Its external surface is of white color, and is in contact with the inner surface of the fascia of the bulb; it is quite smooth, except at the points where the Recti and Obliqui are inserted into it; its anterior part is covered by the conjunctival membrane. Its inner surface is brown in color and marked by grooves, in which the ciliary nerves and vessels are lodged; it is separated from the outer surface of the choroid by an extensive lymph space (spatium perichorioideale) which is traversed by an exceedingly fine cellular tissue, the lamina suprachorioidea.

(a) A passage about "*The Tunics of the Eye*" in Henry Gray's anatomical textbook (Gray, 1918)

- The sclera has received its name from its extreme density and hardness;
- The sclera is a firm, unyielding membrane,
- The sclera is serving to maintain the form of the bulb.
- The sclera is much thicker behind than in front;
- the thickness of The sclera's posterior part is 1 mm.
- The sclera's external surface is of white color,
- The sclera's external surface is in contact with the inner surface of the fascia of the bulb;
- The sclera's external surface is quite smooth, except at the points where the Recti and Obliqui are inserted into The sclera;
- The sclera's anterior part is covered by the conjunctival membrane.
- The sclera's inner surface is brown in color
- The sclera's inner surface is marked by grooves, in which the ciliary nerves and vessels are lodged;
- The sclera's inner surface is separated from the outer surface of the choroid by an extensive lymph space (spatium perichorioideale)
- spatium perichorioideale is traversed by the lamina suprachorioidea.
- the lamina suprachorioidea is an exceedingly fine cellular tissue

(b) Itemized text

Figure 1: We propose to change the prose style of anatomical textbooks, by semi-automatically transforming (a) cohesive text into (b) itemized text, for subsequent information extraction.

interdependence of statements in cohesive text , we propose to change the prose style of anatomical textbooks by semi-automatically transforming cohesive discourse into itemized text – a set of independent statements like the one illustrated in Figure 1(b), obtained by annotating reference expressions and coordinating conjunctions so as to decompose their cohesion.

We then validate the utility of the transformed text for the identification of anatomical NEs and their relations, verifying the assumption that textual cohesion is a major obstacle to accessing the valuable knowledge in textbooks by means other than reading. Specifically, we show that, compared to the original texts, transformed ones are easier to process by machine and hence serve as a convenient way of identifying mentions of anatomical NEs as well as their relations, specifically triple subject–predicate–object relations. Since transformed text is human readable as well, we propose a new model of language resources accessible by both human and machine, improving the computational reusability of anatomical textbooks at less cost to the human reader.

2 RELATED WORK

Over the past few decades and increasingly in recent years, a vast amount of research has been published on information extraction (IE) from unstructured text. A series of Message Understanding Conferences (MUCs) beginning in the 1980s has given focus to this research. IE researchers have been aiming in particular to extract structural information on relations between entities, for example to identify perpetrator, instrument, and target in an incident of terrorism (MUC-4, 1992); successful results have been reported in many domains, including opinion mining (Abe et al., 2011) and Influenza detection (Aramaki et al., 2011). More recently, the traditional IE approach, which pre-specifies relations in the domain in question, has given way to a new paradigm that aims to extract arbitrary, open-domain relations in order to scale to the size of the web for covering a large diversity of relations (Etzioni et al., 2008; Riedel et al., 2013), or to generate LOD for the semantic web (Augenstein et al., 2012). However, both traditional and new IE approaches focus on extracting relations from news or weblog text, which are considered less cohe-

sive (and hence easier to read) than scientific texts.[1] To our knowledge, however, no study has reported IE results from highly cohesive texts like anatomical textbooks. The objective of this paper is not to apply standard IE methodologies to cohesive text, but to demonstrate the effect of cohesion decomposition on IE using anatomical textbooks.

Nevertheless, the set of itemized statements that we transform from out of the textbooks can be considered to constitute a corpus annotated for future automation of anaphora resolution and coordination disambiguation. There are, in fact, already many publicly available corpora annotated for anaphora (see (Ng, 2010)), and a few for coordination such as GENIA (Kim et al., 2003). However these corpora were constructed mainly in order to compile training data for supervised machine learning. To our knowledge, no previous research on corpus annotation has investigated the effects of anaphora resolution and coordination disambiguation on the identification of complex NEs like anatomical terms and of their relations.

3 MOTIVATING EXAMPLES

3.1 Complicated Mentions of Anatomical Named Entities

Many anatomical named entities (NEs) comprise multiple words. For example, in FMA (*The Foundational Model of Anatomy*) (Rosse and Mejino, 2008), a reference ontology in the domain of anatomy, the average number of words that an NE consists of is 6.2 over the total number of 78977 terms.[2]

Thus, mentions of anatomical NEs tend to take the form of long noun phrases consisting of a proper noun and its modifiers, such as prepositional phrases and adjectives. Owing to this, the identification of anatomical NEs is adversely affected by cohesive ties. Because it is tedious for human writers to repeat long terms such as anatomical NEs and for human readers to read them, they are likely to be replaced by reference expressions upon mentions after the first. For example, if the *lateral sulcus* was previously mentioned, an anatomical NE consisting of six words,

> *anterior horizontal limb of the lateral sulcus*

might often be given as

> *anterior horizontal limb of the sulcus*,

where *lateral sulcus* is replaced with its anaphor *the sulcus*. The resulting term will no longer be identifiable by dictionary look-up.

Moreover, when two NEs share words, like

> *anterior horizontal limb of lateral sulcus* and *anterior ascending limb of lateral sulcus*,

a coordinating conjunction may be used to reduce redundancy, as for instance

> *anterior horizontal and ascending limbs of lateral sulcus*.

As indicated above, this manner of writing anatomical terms prevents the identification of mentions of these NEs by exact-match look-up in the dictionary.

3.2 Complicated Sentences

In anatomical textbooks, cohesive ties not only occur frequently but also make up complex reference chains and nested coordinations. For example, the two anaphors (bolded) in the next sentence are chained:

> The thalami are two large ovoid masses. ***The anterior extremity*** is narrow. ***It*** lies close to the middle line... (Gray, 1918)

Because *It* refers to *The anterior extremity* and *The anterior extremity* refers to *The anterior extremity of the thalami*,[3] by summing up the two effects, we see that *It* actually refers to *The anterior extremity of the thalami*.

Another example is the following sentence containing two coordinated phrases, induced by conjunction *and*, that are nested:

> The posterior extremity is expanded, is directed backward *and* lateralward, *and* overlaps the superior colliculus. (Gray, 1918)

Nested coordination makes a sentence more complex and longer, and thus more difficult to parse (Hara et al., 2009). This causes a problem, namely the identification of relations as subject–predicate–object structures on the basis of erroneous parser outputs.

[1] For example, the percentage of sentences that contain coordinating conjunction (such as "*and*" and "*or*"), tagged as "CC," is 44.9% for news texts in the *Wall Street Journal* part of the Penn Treebank (Marcus et al., 1993), but 58.8% for biomedical articles in the GENIA corpus (Kim et al., 2003), although average sentence length is almost the same in these corpora (23.9 words for the Penn Treebank and 23.4 words for GENIA).

[2] One reason for this is that anatomical NEs likely contain prepositional phrases. For instance, the percentage of terms in FMA that contain the preposition "*of*" is 78.3%.

[3] In Figure 2, instead of considering that *The anterior extremity* refers to *The anterior extremity of the thalami*, we suppose for convenience that *The* refers possessively to *Thalami*.

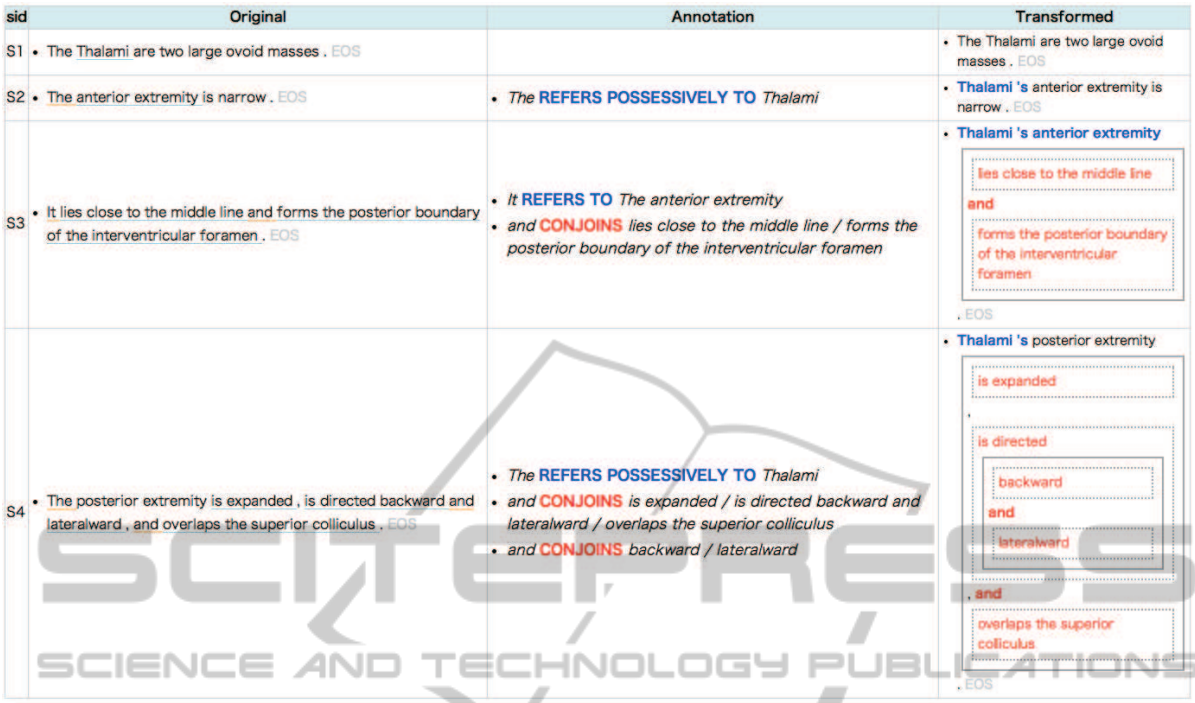

Figure 2: A web GUI we developed for the annotation of reference expressions and coordinating conjunctions. The first and second columns denote sentence IDs and original texts. The third column, which is empty at the beginning, shows annotations added by a worker. The fourth column shows texts transformed according to the annotations.

4 DEMONSTRATION

We now transform an anatomical textbook into itemized text; then, we validate the utility of the transformed text for the identification of anatomical NEs and their relations.

We use "*Anatomy of the Human Body*", a textbook written by Henry Gray (Gray, 1918). We select this textbook because although the anatomical knowledge presented in it was mostly established in earlier days and is now in the public domain, removing licensing issues, it is nevertheless mostly current.

Gray's textbook comprehensively describes the morphological features of and relations among human body parts, such as the bones, muscles, nerves, and so on. We focused on the "*Fore-brain or Prosencephalon*" section,[4] which describes the most complicated structure in the human body.

We split the text manually and obtained 787 test sentences. The percentage of sentences that contain coordinating conjunction "*and*" is 56.8%, and average sentence length is 23.2 words.

[4]http://www.bartleby.com/107/189.html

4.1 Semi-automated Itemization

4.1.1 Manual Annotation.

Using the web GUI shown in Figure 2, Gray's text is annotated to allow the decomposition of the cohesion induced by reference expressions and coordinated conjunctions. Annotations are done by a worker using a computer mouse to select from the original texts two or more sequences of words (e.g., an anaphor and its antecedent, or a coordinating conjunction and its conjoined elements) and link them with a label, according to the two tasks below:

Anaphora Resolution. For each anaphor (whether an anaphoric pronoun or an anaphoric noun phrase) in the original texts, select its antecedent, and connect anaphor and antecedent with a label "REFERS TO." In the GUI, this annotation is denoted as

anaphor REFERS TO *antecedent*.

Note that for indirect anaphora in which a possessive pronoun or anaphoric determiner is involved, a different label is selected, "REFERS POSSESSIVELY TO."

The Thalami are two large ovoid masses.
Thalami's anterior extremity is narrow.
Thalami's anterior extremity lies close to the middle line.
Thalami's anterior extremity forms the posterior boundary of the interventricular foramen.
Thalami's posterior extremity is expanded.
Thalami's posterior extremity is directed backward.
Thalami's posterior extremity is directed lateralward.
Thalami's posterior extremity overlaps the superior colliculus.

(a) Transformed text

< Thalami's anterior extremity, form, posterior boundary of interventricular foramen >
< Thalami's posterior extremity, overlap, superior colliculus >

(b) Relations (<subject, predicate, object>) extracted from transformed text

Figure 3: Example of relation extraction from transformed texts (using the same texts as in Figure 2.)

Coordination Disambiguation. For each occurrence of a coordinating conjunctions such as *and*, find a series of elements (verbal phrases) that are conjoined by the conjunction, and mark the whole with a label "CONJOINS." In the GUI, the annotation is denoted as

coordinating conjunction CONJOINS *a series of elements*.

Across the 787 test sentences, we attached 507 "REFERS TO" labels, 165 "REFERS POSSESSIVELY TO" labels, and 783 "CONJOINS" labels.

4.1.2 Transformation

When text is annotated, transformation of the original text can be performed automatically according to the annotation. With reference to the "REFERS TO" labels, original texts are transformed by replacing the anaphors with their (possibly chained) antecedents (or with an apostrophe plus the letter s for the "REFERS POSSESSIVELY TO" labels). For an original sentence to which was attached a "CONJOINS" label, the sentence are duplicated and each of the conjoined elements is distributed. For example, the third and fourth sentences in Figure 3(a) are generated by duplicating the third sentence of the original text in Figure 2, which contains one coordinated phrase.

Ultimately, we transformed the original 787 sentences into a set of itemized statements constituting 1871 context-independent sentences.

4.2 Information Extraction

Now we compare the original and the transformed text with respect to the results for the identification of mentions of anatomical NEs and their subsequent relation extraction.

Table 1: The number of mentions of anatomical NEs identified from 787 test sentences, before (= original text) and after (= transformed text) decomposing cohesion.

	# anatomical NEs	diff.
Original text	1641	—
Transformed text	**2194**	+553

Mention Identification. To identify the anatomical NEs mentioned in the text, we looked them up in FMA,[5] a reference ontology of anatomy (Rosse and Mejino, 2008). Table 1 shows the number of identified mentions of anatomical NEs, and indicates that this number was increased by the decomposition of cohesion.[6] This is partly because the 15 anatomical NEs listed in Table 2 in the transformed text that are missing in the original text are discovered by dictionary look-up.

Relation Extraction. Next, we employed Enju,[7] a state-of-the-art natural language parser, and applied it to the text. Specifically, we parsed the original and the transformed text, and from among the subject–predicate–object structures output by Enju, we picked up those that contained anatomical NEs (identified by the previous process of mention identification) in both the subject and the object. Table 3 shows the result. The number of extracted relations, as well as the number of those that were anatomically correct,[8] was greatly increased by transformation. For example, we

[5] http://www.berkeleybop.org/ontologies/fma.obo

[6] We do not count duplicate mentions of NEs on the basis of the "CONJOINS" labels.

[7] http://www.nactem.ac.uk/enju/

[8] The correctness was judged by a medical doctor.

Table 2: Anatomical NEs that are missing in the original text but are discovered in the transformed text by dictionary look-up.

FMA ID	Anatomical Term Name
50087	Anterior choroidal artery
50655	Calcarine artery
52573	Inferior branch of oculomotor nerve
59669	Roof of internal nose
61944	Anterior forceps of corpus callosum
62418	Lateral orbital gyrus
67956	Medial longitudinal stria
83759	Anterior ascending limb of lateral sulcus
83760	Anterior horizontal limb of lateral sulcus
83761	Posterior ascending limb of lateral sulcus
84114	Apical part of cell
256305	Lateral surface of cerebral hemisphere
256312	Basal surface of cerebral hemisphere
256318	Medial surface of cerebral hemisphere
256335	Tentorial surface of cerebral hemisphere

Table 3: The number of relations (subject–predicate–object triples that contain anatomical NEs in both subject and object) extracted from 787 test sentences, and the number of those that were anatomically correct, before (= original text) and after (= transformed text) decomposing cohesion.

	# triples	# correct
Original text	70	45
Transformed text	**366**	**310**

can extract the relations shown in Figure 3(b) from the transformed texts, but not from the original ones.

5 CONCLUSION

In this paper, we proposed to transform the prose style of anatomical textbooks by annotating reference expressions and coordinating conjunctions. We then validated the utility of the transformed text for the identification of anatomical NEs and their relations, and verified that the cohesiveness of the text is one of the major obstacles preventing us from accessing knowledge in textbooks by methods other than reading.

Since the transformed text is human readable as well, the proposed style has potential to serve as a new-model language resource that is accessible by both human and machine, promising to improve the computational reusability of anatomy textbooks. We also believe the proposed method to be applicable to texts in domains other than anatomy, as long as they mainly consist of factual explanation of structures, for instance natural or artificial geographical and geological features.

REFERENCES

Abe, S., Inui, K., Hara, K., Morita, H., Sao, C., Eguchi, M., Sumida, A., Murakami, K., and Matsuyoshi, S. (2011). Mining personal experiences and opinions from web documents. *Web Intelligence and Agent Systems*, 9(2):109–121.

Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: Detecting influenza epidemics using twitter. In *EMNLP*, pages 1568–1576.

Augenstein, I., Padó, S., and Rudolph, S. (2012). Lodifier: Generating linked data from unstructured text. In *Proceedings of the 9th Extended Semantic Web Conference (ESWC)*, pages 210–224.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Commun. ACM*, 51(12):68–74.

Gray, H. (1918). *Anatomy of the Human Body*. Philadelphia: Lea & Febiger, 20 edition.

Hara, K., Shimbo, M., Okuma, H., and Matsumoto, Y. (2009). Coordinate structure analysis with global structural constraints and alignment-based local features. In *ACL-IJCNLP*, pages 967–975, Suntec, Singapore. Association for Computational Linguistics.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). Genia corpus - a semantically annotated corpus for bio-textmining. In *ISMB (Supplement of Bioinformatics)*, pages 180–182.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden. Association for Computational Linguistics.

Riedel, S., Yao, L., McCallum, A., and Marlin, B. M. (2013). Relation extraction with matrix factorization and universal schemas. In *NAACL-HLT*, pages 74–84, Atlanta, Georgia. Association for Computational Linguistics.

Rosse, C. and Mejino, J. L. V. (2008). The Foundational Model of Anatomy Ontology Anatomy Ontologies for Bioinformatics. In Burger, A., Davidson, D., and Baldock, R., editors, *Anatomy Ontologies for Bioinformatics*, volume 6 of *Computational Biology*, chapter 4, pages 59–117. Springer London, London.

Schäfer, U., Spurk, C., and Steffen, J. (2012). A fully coreference-annotated corpus of scholarly papers from the acl anthology. In *COLING (Posters)*, pages 1059–1070.

Witte, S. P. and Faigley, L. (1981). Coherence, cohesion, and writing quality. *College Composition and Communication*, 32:189–204.