Automatic Semantic Subject Indexing of Web Documents in Highly Inflected Languages

Reetta Sinkkilä, Osma Suominen and Eero Hyvönen

Semantic Computing Research Group (SeCo),
Aalto University, Department of Media Technology
University of Helsinki, Department of Computer Science
firstname.lastname@tkk.fi, http://www.seco.tkk.fi/

Abstract. Structured semantic metadata about unstructured web documents can be created using automatic subject indexing methods, avoiding laborious manual indexing. A succesful automatic subject indexing tool for the web should work with texts in multiple languages and be independent of the domain of discourse of the documents and controlled vocabularies. However, analyzing text written in a highly inflected language requires word form normalization that goes beyond rule-based stemming algorithms. We have tested the state-of-the art automatic indexing tool Maui on Finnish texts using three stemming and lemmatization algorithms and tested it with documents and vocabularies of different domains. Both of the lemmatization algorithms we tested performed significantly better than a rule-based stemmer, and the subject indexing quality was found to be comparable to that of human indexers.

1 Introduction

The Semantic Web vision requires structured ontological metadata in order to provide novel services such as rich search interfaces, automatic recommendations, agent-based assistants and semantic personalization. The current Web, however, consists largely of unstructured text documents. Manually annotating such content is often infeasible due to the large amount of work involved, and in any case may not always produce good results [3].

One important method for creating structured descriptions of unstructured text is automatic *subject indexing*, also known as *term assignment*, which is the process of summarizing the content of a document by selecting multiple subjects from a controlled vocabulary that describe its topic [9, 10]. Many automatic subject indexing tools exist for various languages and domains [18]. For example, many systems have been developed for automatically assigning subjects from the Medical Subject Headings (MeSH) vocabulary for biomedical documents [20].

A succesful automatic subject indexing tool for the Web should be flexible enough to work with documents in different languages and domains. The quality of the automatically assigned subjects can usefully be compared with traditional manual subject indexing. However, when two humans describe the same document, they are very unlikely to select the same subjects [24, 8, 16].

Thus, rather than relying on the subjects assigned by a single human indexer as a gold standard, it is more useful to compare the degree of consistency between an automated algorithm and several independent human indexers [10].

Subject indexing algorithms perform language analysis and normalization such as stemming [10, 18]. However, in agglutinative and highly inflected languages such as Finnish [6], Turkish [12], Estonian, Hungarian and Slavic languages, a simple stemming strategy is unlikely to perform well [4, 6].

In this paper, we set out to find a strategy for automatic subject indexing for inflectional languages, using Finnish as the test case with an intention to produce annotations of comparable quality to those produced by human indexers. Finding such a strategy would allow us to substantially increase the amount of structured metadata for use in Finnish Semantic Web portals such as HealthFinland[1] and CultureSampo[2]. In particular, we seek answers to the following research questions:

1. What kind of **stemming** or **lemmatization** strategy gives the best results when performing automatic subject indexing for web documents in highly inflected languages?
2. What is the **quality** of automatically assigned terms for documents written on inflected languages compared with human indexers?
3. Does the same automatic term assignment strategy work **independently of the domain** of the documents and vocabularies?

To answer these questions, we performed a series of automatic subject indexing experiments on Finnish language documents. We used the Maui indexing tool, a language- and domain-independent system which incorporates state-of-the-art topic ranking algorithms and can perform subject indexing using a controlled vocabulary [10]. To test the effect of morphological analysis strategies, we coupled the Maui tool with several available stemming and lemmatization tools. We compared the results of the automatic indexing with subjects assigned by human indexers. To compare the performance of the tools in different domains, we used document sets and vocabularies from two domains: a) documents from a social sector website together with a health-oriented ontology, and b) point of interest descriptions from Wikipedia together with a general purpose ontology.

Our results indicate that the Maui subject indexing algorihms work relatively well even with Finnish language documents when Maui is coupled with a capable lemmatization system, and the indexing quality is comparable to that of human indexers. However, disambiguation between similar or overlapping concepts in the vocabulary was problematic in some cases. Also, the handling of set phrases and compound words caused some issues. Some of the ambiguities might be resolved by using part-of-speech information as an aid in disambiguation. The results should generalize to other highly inflected and agglutinative languages. Even the subject indexing of English documents might be improved by performing more sophisticated linguistic analysis than simple rule-based stemming.

[1] http://www.tervesuomi.fi
[2] http://www.kulttuurisampo.fi

2 Related Work

Automatic subject indexing consists of two phases: performing linguistic analysis for matching document words or n-grams with meanings expressed as terms in a controlled vocabulary (*semantic tagging*) and determining which of the matched vocabulary terms best describe the document (*topic ranking*).

Semantic tagging is the matching of words to meanings and a part of linguistic analysis. Linguistic analysis for the purpose of annotation consists of five steps: morphological analysis, part-of-speech tagging, chunking, dependency structure analysis and semantic tagging [1]. In languages such as English, Spanish and French, a simplified form of semantic tagging can be performed by using a rule-based stemming algorithm to normalize both document words and vocabulary terms [10]. This allows, e.g., singular words to be matched with plural terms in the vocabulary.

Inflected languages such as Finnish, Turkish, Arabic and Hungarian typically express meanings through morphological affixation. In highly inflected languages plural and possessive relations, grammatical cases, and verb tenses and aspects, which in English would be expressed with syntactic structures, are characteristically represented with case endings [12, 6]. Compound words are also typical in inflected languages. Rule-based stemming does not work particularly well for semantic tagging: as an example, a semantic tagger for the Finnish language developed in the Benedict project used a sophisticated morphological analysis and lemmatisation tool as well as rules for handling compound words in order to attain high precision [6, 7].

In **topic ranking**, machine learning methods have surpassed rule-based methods for determining the important topics of a document [18]. The TF×IDF method provides a baseline [17], which Maui [10] and its predecessors KEA [23] and KEA++ [11] have improved on by additionally using various heuristics. These tools can also perform topic indexing without the support of a controlled vocabulary, known as *keyphrase extraction*. The previous Maui tests on English, French and Spanish docments have used a stemming algorithm for basic semantic tagging. In those languages, Maui has been found to assign subjects of comparable quality of those of human indexers [10].

KEA has been ported to support other languages. A Turkish adaptation of KEA was used to extract keyphrases and a controlled vocabulary was not used [13]. A KEA-like approach for keyphrase extraction of Arabic documents has also been found to perform well when part-of-speech analysis was incorporated into the candidate selection phase [2].

Other subject indexing tools for inflected languages include the Poka information extraction tool for Finnish [21], which has been used, e.g., in the Opas system to assign concepts from the Finnish General Upper Ontology to question-answer pairs [22]. The Leiki platform is a commercial tool that analyzes Finnish text and determines its important concepts using a proprietary ontology-like classification system [14]. It is used by many Finnish news websites for automatically generating links to related content. However, neither tool has been evaluated in academic literature.

3 Materials and Methods

For our experiments, we have used three document collections together with two vocabularies. With these, we performed three experiments using the Maui indexer and three different stemming and lemmatization tools.

3.1 Document Collections

To provide material for experiments we prepared two corpora and annotated them with different vocabularies. This was to ensure that we can measure the performance of the automatic indexing independently of the domain of the documents and vocabularies.

The first text corpus consists of documents extracted from the Sosiaaliportti web portal[3] maintained by the Finnish National Institute for Health and Welfare. Sosiaaliportti is designed for professionals in the social sector, and is intended to support social workers in their daily work. It contains 1) question-answer pairs on topics related to social work in general, such as "What are the criteria for granting a transportation service for a severly disabled person", 2) a discussion forum, 3) the handbook for child welfare which is intended to be used as a topical manual for professionals.

The first Sosiaaliportti document collection we used, **SOS-60**, consists of 60 randomly extracted documents from the Sosiaaliportti portal. This sample includes 30 documents from the Handbook for child welfare and 30 question-answer pairs from the consultancy service archive. The documents are relatively short, ranging from 33 to 1324 words with an average of 360 words.

The second document collection **SOS-30** is a subset of SOS-60, consisting of 15 question-answer pairs and 15 Handbook documents. It was created in order to determine the inter-indexer consistency of human indexers.

The document collections were indexed by employees of the National Institute for Health and Welfare – professionals ranging from a summer trainee to a medical doctor. Indexers were advised to use 3–8 subjects per document, which is the usual amount of index terms used in the National Institute for Health and Welfare content indexing process. The SOS-60 collection was indexed by a single person, who assigned an average of 5 subjects per document. The SOS-30 collection was indexed by six people, with an average of 4.1 subjects per document. The mimum number of assigned subjects was 0 (two indexers used this) and maximum number was 9. Summary of the number of subjects used by indexers is in table 1. Both datasets were created and indexed for the purpose of the experiments reported in this paper.

To test the domain independence of the Maui topic extractor, another document collection **POI-61** was created, consisting of 61 documents extracted from the Finnish Wikipedia with subjects covering Finnish Points of Interest (POIs) such as churches and statues. Characteristically these documents are also relatively brief, containing 450 words per document on the average. The POI-61

[3] http://www.sosiaaliportti.fi

Table 1. Number of subjects assigned to each document in SOS-30

	Min	Max	Mean	St. deviation
Indexer 1	0	6	2.8	1.6
Indexer 2	1	9	3.7	1.7
Indexer 3	0	8	4.6	1.7
Indexer 4	3	6	3.9	0.8
Indexer 5	2	8	4.4	1.6
Indexer 6	2	8	5.3	1.5
Average	1.3	7.5	4.1	1.5

collection was indexed by a single person. The average number of subjects per document was 7.6 with a minimum of 1 and maximum of 15 subjects.

The charasteristics of the document collections are slightly different. The Sosiaaliportti collections consist of shorter documents that are indexed with fever terms per document, while the Wikipedia corpus is more exhaustively annotated. Also the content and structure of the documents differs. Sosiaaliportti documents have been written by professionals and they cover topics more in-depth. The Wikipedia documents are more of a descriptive nature.

3.2 Vocabularies

The Sosiaaliportti document collections were indexed using concepts from the Finnish Ontology of Health and Welfare, Tero. It is a combination of several health domain vocabularies including The European Multilingual Thesaurus on Health Promotion (HPMULTI)[4] and a subset the Medical Subject Headings (MeSH) thesaurus[5], merged with the Finnish General Upper Ontology YSO[6]. YSO is based on the Finnish General Thesaurus maintained by the National Library of Finland.

Tero defines over 24 000 concepts which have Finnish, Swedish and English labels. Only the Finnish labels have been used in the indexing process described in this research. There are also alternative labels for some concepts such that the total amount of Finnish labels in the ontology is around 30 000. Tero is represented in SKOS format and the relations between terms have been represented according to SKOS conventions, e.g. the skos:broader relation representing a hierarchical relation. Tero contains some ambiguous terms with disambiguating context information coded in parenthesis, for example *lapset (perheenjäsenet)* and *lapset (ikäryhmä)* standing for *children (family members)*, and *children (age group)*, respectively. Some of these ambiguous concepts have the unqualified ambiguous labels as alternative label, e.g. *lapset*.

[4] http://www.hpmulti.net/

[5] http://www.nlm.nih.gov/mesh/

[6] http://www.seco.tkk.fi/ontologies/yso/

The POI-61 documents were indexed using the Finnish Collaborative Holistic Ontology, Koko. It is a collection of Finnish core ontologies that have been merged together. The ontologies include the Finnish General Upper Ontology YSO as its top ontology and a variety of other domain specific ontologies extending its concepts into more detailed subconcept hierarchies. These include for example the ontology for museum domain, the ontology for applied arts and the Finnish ontology for photography. Koko defines some 30 000 concepts and is encoded in SKOS format. Concepts have preferred and alternative labels in Finnish, Swedish and English. Only the Finnish labels were utilized in these experiments.

Table 2. Sizes of the Tero and Koko vocabularies

	Total terms	PrefLabels	AltLabels	Ambiguous
Finnish Tero	30,040	24,270	5,770	1720
Finnish Koko	38,690	30,080	7,810	1910
English Agrovoc	38,200	28,170	10,030	400
French Agrovoc	37,350	28,160	9,190	440
Spanish Agrovoc	40,640	28,160	12,480	620

A summary of vocabularies used in this research is shown in table 2. For comparison, the table also includes the corresponding statistics of the Agrovoc thesaurus which was used in the original Maui experiments [10]. Both Tero and Koko contain a relatively large number of ambiguous terms compared to Agrovoc. This is due to the inclusion of the Finnish Upper General Ontology, which contains a large amount of everyday terms which more often have several meanings than domain-specific specialist terminology such as Agrovoc terms.

3.3 Maui Topic Indexing Tool

We selected the Maui topic indexing tool, version 1.2, for our automatic indexing experiments as it implements a state-of-the art topic ranking algorithm [10]. Although Maui can be used without a controlled vocabulary, we will concentrate on the case when a vocabulary is used. The topic ranking is based on a number of heuristics (called *features* in Maui terminology) including TF×IDF, spread (separation of first and last occurrence), semantic relatedness based on vocabulary structure, and term length. The algorithm is first tuned with a small training set of manually indexed documents, which is used to tune the relative weights of the heuristics. After training has been completed, it can perform subject indexing on new documents.

In the indexing phase, Maui first splits the text into textual segments (usually sentences). These are then further split into words, which are then grouped into n-grams. The n-grams are matched with terms in a controlled vocabulary; stemming is performed both on the n-grams and the vocabulary terms in order to increase recall, for example by matching singular form words with plural forms in the vocabulary. The stemming algorithms are language specific, but new stemmers can be plugged in.

Finally, the n-grams which were determined to match vocabulary terms are ranked by applying the different heuristics and summarizing the feature values according to the weights that were determined in the training phase. The top K matched vocabulary terms are assigned as subjects for the document.

3.4 Stemming and Lemmatization Tools

We inspected the effect of different word form normalizations by testing three methods for deriving base forms of words. We tested the commercial syntactic dependency parser FDG version 3.8.1 for Finnish [19] by Connexor Ltd, the morphological analyzer Omorfi version 20100401 for Finnish [5] and the Snowball stemmer for Finnish[7].

The main difference between these tools is that while FDG and Omorfi try to reduce the word forms into their lemmas base forms, the Snowball stemmer only stems word by cutting off inflectional suffixes without fixing the consonant gradiation. Another difference between the selected tools is how they handle compound words and set phrases as well as words unfamiliar to them. For example, the word *seurakuntatyö* (*church/parish work*) is a coumpound word consisting of parts *seurakunta* and *työ*, with meanings *parish* and *work*, respectively. The word *seurakunta* could also be split into its constituent parts *seura* and *kunta*, but the compound word has a special lexicalized meaning which does not directly follow from the parts. The FDG parser lemmatizes the word correctly recognizing the fixed compound semantic meaning and handles the word as a compound word. The version of the Omorfi parser we used instead returns every possible combination of the word parts without any weights indicating some interpretation as more favorable. With unfamiliar words Omorfi returns the original input, whereas FDG and Snowball always try to reduce the word to a base form.

3.5 Experiments

We conducted three sets of experiments in order to answer our research questions. The purpose of the first experiment was to determine the effect of lemmatization method used. The second experiment compares the quality of automatically assigned terms with human indexing. The last test set tries to evaluate whether the automatic term assignment strategy works independently of the domain of the documents and vocabularies.

Stemming and lemmatization strategy The first experiment was to test how well Maui performs with Finnish data using different stemming and lemmatization strategies. We used the SOS-60 document collection and the Tero vocabulary for this experiment. To provide a useful point of comparison, the experiment setup closely followed the experiment described in [10], section 7.2.4.

[7] http://snowball.tartarus.org/

Language independence, a test conducted with collections of 67 French and 47 Spanish agricultural documents indexed with Agrovoc thesaurus terms.

The test settings were the following: the stemmer was set to either FDG, Omorfi or Snowball. The stopword list was set to a list of Finnish stopwords taken from the Snowball string processing language site[8]. Document encoding was set to UTF-8. Document language was set to *fi*, to use the Finnish labels of the vocabulary. Tests were conducted with the leave one out technique. That is, the maximum possible number of documents (59) were used to create a model and the one remaining document was used for testing the model. The tests were repeated 60 times, each time indexing a different document. This is the same approach which was taken in the original tests with Spanish and French documents [10]. For each document 5 terms were extracted, which was the average number of manually assigned subjects per document in SOS-60. We re-ran the tests using each stemming or lemmatization tool in turn.

In addition, to test the effect of the Maui topic ranking algorithms, we used the term frequency – inverse document frequency method TF×IDF as a baseline by turning off all other Maui topic ranking algorithms. It was calculated only using terms from the vocabulary as candidates, and FDG as a lemmatizer.

Inter-indexer consistency In the second experiment, we tested how well Maui performs related to human indexers. We used the SOS-30 document collection, indexed by six independent people, and the Tero vocabulary. To provide a reference for evaluation, we first measured the performance of the independent human annotators by measuring the similarity between indexers. We measured the consistency with the Rolling measure [15], defined as

$$\frac{2C}{A+B} \tag{1}$$

where C is the amount of subjects two indexers have in common and A and B the amount of subjects used by indexer A and B respectively. With this measure, two identically annotated documents get a similarity value of 1 and totally distinct annotations get a value of 0. We counted this measure for each document between every indexer-indexer pair. The total consistency between two indexers is the average of the document specific values.

We then indexed the same document set with the Maui topic extraction tool using each human annotated document set as training material in turns. Similarly to the first experiment, this was done with the leave-one-out technique to maximize the available training material. For each document 4 subjects were assigned, which was the average number of terms the human indexers had assigned per document in SOS-30. We used FDG to perform lemmatization in this experiment. The other parameters were the same as in the first experiment.

Automatic annotations were then compared to the manually made annotations, first in pairs with each annotator and finally the average agreement was calculated with the same procedure as that used between human indexers. This

[8] http://snowball.tartarus.org/algorithms/finnish/stop.txt

made it possible to directly compare the performance of Maui to a human indexer.

Domain independence To test the domain independence and suitability of the method for different materials, we conducted a third experiment with a different document collection and vocabulary. The POI-61 document collection and KOKO vocabulary were used in this experiment. The test was conducted in a similar way to the first experiment, but using only FDG for lemmatization.

4 Results

In this section, we present the results of the three experiments described above.

4.1 Stemming and Lemmatization Strategy

The first test setting was to test the suitability of Maui tool for Finnish language with alternating stemmers. The Maui topic extractor was ran with vocabulary language set to Finnish and stemmers set to FDG, Omorfi and Snowball in turns. We used the SOS-60 collection which is indexed with the TERO ontology.

Table 3. Stemming and lemmatization strategy results

	Precision	Recall	F-Measure
SOS-60, FDG	40.0	37.1	38.5
SOS-60, Omorfi	40.0	35.9	37.8
SOS-60, Snowball	35.7	32.2	33.8
SOS-60, FDG, TF×IDF only	27.0	24.4	26.7
French Agrovoc	34.5	31.8	33.1
Spanish Agrovoc	24.7	26.9	25.7

Compared to the baseline method, TF×IDF, the Maui topic extraction algorithm performed better regardless of which parser was used (Table 3). For comparison, the figures from the Maui tests with French and Spanish documents [10] are also included in the table. The best lemmatisation strategy was FDG but Omorfi also showed good results. The precision was the same for both strategies, but the ones performed with FDG resulted in better recall.

Two documents indexed with different methods are shown in Table 4. For each document terms assigned by a professional indexer and by Maui tool with FDG, Snowball and Omorfi parser, respectively, are presented. The first document is an example of a succesful automatic annotation. Annotations made with Maui tool with all of the lemmatisation strategies performed well compared to the human indexer. Correct terms are emphasized, inapplicable or redundant terms are marked with cursive text.

The second document is an example of an unsuccesful annotation procedure with some peculiarities. The human indexer has assigned topics related to Romani people, clothes and pregnancy. Automatically assigned terms include term *wood chip*, which is inapplicable with regards to the document's contents. This is a result of imperfect morphological analysis. That is, the document's contents have been processed with a stemmer, which has produced a stem, which has in turn been connected to a different term with the same stem. Also, some identical topics have been chosen to describe the same document. Maui tool has assigned terms *nuoret* (*young people*) and *nuori (13-18)* (*adolescent*), with overlapping meaning, the only difference being the first one in plural form and the second one accompanied by the specification *13-18*.

Table 4. Example documents

	Document 1 (good performance)		Document 2 (bad performance)	
	Finnish	English	Finnish	English
Human Indexer	perhehoito	foster care	vaatteet	clothes
	kiireellinen sijoitus	high-priority placement	raskaus	pregnancy
	huostaanotto	placement into care	romanit	Romani people
	avohuollon tukitoimet	support in community care	romanit - kulttuuri	Romani culture
	jälkihuolto	after-care		
	laitoshoito	institutional care		
	sijaishuolto	substitute care		
Maui + FDG	**sijaishuolto**	**substitute care**	*nuori (13-18)*	*adolescent*
	avohuollon tukitoimet	**support in community care**	ohjeet	instructions
	jälkihuolto	**after-care**	hakemukset	applications
	lapset (sosioek.)	children (socioeconomic)	**raskaus**	**pregnancy**
	huostaanotto	**placement into care**	*nuoret*	*young people*
Maui + Omorfi	**sijaishuolto**	**substitute care**	synnytys	delivery
	lapset (sosioek.)	children (socioeconomic)	**raskaus**	**pregnancy**
	jälkihuolto	**after-care**	nuori (13-18)	adolescent
	huostaanotto	**placement into care**	ohjeet	instructions
	avohuollon tukitoimet	**support in community care**	*kaulukset*	*collar*
Maui + Snowball	**sijaishuolto**	**substitute care**	nuori (13-18)	adolescent
	lapset (sosioek.)	*children (socioeconomic)*	hakemukset	applications
	jälkihuolto	**after-care**	naiset	women
	avohuollon tukitoimet	**support in community care**	*hake*	*wood chip*
	lapsi (6-12)	*child (6-12)*	*kunnat*	*municipalities*

4.2 Inter-Indexer Consistency

The results of the inter-indexer consistency experiment are shown in Table 5 using the Rolling measure. The consistency between any two indexers can be found in the table. The average consistency of the human indexers was 33.7%.

This corresponds to 22.6% on Hooper's scale [24]. Previous studies have reported inter-indexer consistencies to vary widely subject to, e.g., the previous experience of the indexers and the usage of controlled vocabulary [8]. [16] reports consistency close to 20%, while [8] presents consistencies between 10–80%.

There is some variance between the indexers. Indexer 1 is least consistent with the other human indexers (27.4%) while Indexer 5 agrees the most with the other indexers (36.6%).

Table 5. Consistency of human indexers 1–6 compared to Maui

	1	2	3	4	5	6	Average	**Maui**
1		25	29	28	27	28	27.4	**21.5**
2	25		31	30	36	37	31.8	**29.9**
3	29	31		40	42	39	36.2	**27.2**
4	28	30	40		38	35	34.2	**36.3**
5	27	36	42	38		40	36.6	**25.3**
6	28	37	39	35	40		35.8	**27.2**
							33.7	**27.9**

The Maui topic indexing algorithm is 27.9% consistent with human indexers. Maui indexes most alike with the Indexer 4 (36.3%) and least alike the with Indexer 1 (21.5%). There is quite a lot of variance between the performance of the automatic annotation method when compared with different indexers. The Maui indexer acts poorly with Indexer 1's document collection as training material. This might result from Indexer 1 using fewer than three terms per document, the average being four terms.

4.3 Domain Independence

The results of the third experiment using POI-61 and Koko were similar to those of the first SOS-60 collection test (Table 6), with slightly higher precision and recall values attained.

Table 6. Domain independence experiment results

	Precision	Recall	F-Measure
SOS-60 with FDG	40.0	37.1	38.5
POI-61 with FDG	45.4	38.1	41.4

5 Discussion and Future Work

In this section, we revisit the research questions based on the results, highlight some problems we encountered and present opportunities for future work.

What kind of stemming or lemmatization strategy gives the best results when performing automatic subject indexing for web documents in highly inflected languages? Our first experiment with three different stemming and lemmatization methods demonstrated that both Omorfi and FDG can be used for lemmatization and both will give good results. The simple rule-based Snowball stemming algorithm did not work as well.

The quality of indexing using Omorfi was almost as good as with FDG. This may at first be surprising, because Omorfi only analyzes the morphological structure of individual words, which may be ambiguous. In contrast, FDG is able to perform part-of-speech and dependency structure analysis. However, in this case the way Maui chunks sentences into individual words before stemming prevents FDG from seeing the whole sentence, thus making it impossible for FDG to analyze the word context. This presents an opportunity for future work: if Maui were adapted so that it is able to pass full sentences to the stemming algorithm, better indexing quality might be attained when using a more sophisticated lemmatization algorithm such as FDG. This kind of experiments could also be performed with less inflected languages such as English.

What is the quality of automatically assigned terms for documents written on inflected languages compared with human indexers? The inter indexer consistency test found consistency between indexers to be 33.7%, whereas consistency between Maui and the human indexers was 27.9%. Maui annotates topics almost as well as human indexers, but there are rather large differences between indexers. The performance of Maui in terms of agreement with human indexers was slightly higher than that of Indexer 1, who had the lowest agreement score (27.4%). The result is somewhat better than in a previous similar evaluation, where the performance of Maui was lower than that of every professional human indexer [10, Table 7.7].

Human indexers are notoriously unreliable when unmotivated, for example taking shortcuts when asked to perform topical indexing as part of a publication process [3]. In our second experiment, some indexers used much fewer subjects per document than they were asked for, and left some documents unindexed. An automated algorithm may not perform as well as motivated professional indexers, but its results can be expected to be more consistent with the task specification.

Does the same automatic term assignment strategy work independently of the domain of the documents and vocabularies? It has previously been shown that the Maui algorithm works with documents and vocabularies of different domains, including the medical, physics and agriculture domains

[10]. The results of our third experiment using point of interest descriptions and a general ontology suggest that when a suitable lemmatizer is used the algorithm also works well with Finnish text of different domains.

Problems Encountered The most essential problem we encountered with the topic indexing with the selected methods, were related to disambiguation of the vocabulary terms. The Maui tool sometimes selected overlapping topics (Table 4) and was not always able to disambiguate between different meanings even though a semantically linked vocabulary was available. Especially problems arise when concepts share equal labels, which often happens with alternative labels in general purpose ontologies. This issue was especially pronounced when stemming was used instead of more sophisticated morphological analysis.

Further problems arose with set phrases, which the Maui tool can not handle as a unit. If the document contains terms *tunnustelu* (examination) and *käsi* (hand), Maui may assign it to a compound term from the vocabulary *käsin tunnustelu* (examination with hands).

Some disambigation problems might be avoided if the words of the documents were not considered in a bag of words style, where the possibility to disambiguate words based on part of speech or dependency structure is lost. If documents were sent to a syntactic dependency parser sentence by sentence, then some misinterpretations could be avoided.

Future Work There is still room for improvement in topic indexing for inflective languages, particularly by using sentence-level analysis and part-of-speech disambiguation as discussed above. We are also looking at ways to simplify the use of information extraction methods for the automatic annotation of text documents that can then be incorporated into Semantic Web portals. We have produced an initial prototype of ARPA, which is an automatic annotation system that provides API access similar to the OpenCalais toolkit[9]. When completed, ARPA will feature a subject indexing facility based on Maui as well as an ontology-based named entity recognition facility.

6 Conclusion

A good automatic subject indexing algorithm makes it possible to substantially increase the amount of structured metadata on the Semantic Web. However, most research to date has concentrated on English language documents, where the language analysis can be performed by a simple rule-based stemmer.

Automatic subject indexing using stemmers is difficult in inflective languages such as Finnish, Turkish, Arabic and Slavic languages. In our experiments on Finnish documents and vocabularies using the Maui indexing toolkit, we were able to increase indexing quality by using more sophisticated lemmatization algorithms Omorfi and FDG instead of a simple rule-based stemmer. Using similar

[9] http://www.opencalais.com/

analysis tools would be useful for subject indexing other heavily inflected languages.

The subject indexing quality we attained was comparable to that of human indexers, in line with earlier similar experiments on documents in other languages. Indexing quality might yet be improved by using part-of-speech information and dependency structure analysis in the semantic tagging phase. Also, such a strategy might assist in disambiguating between similar concepts in a controlled vocabulary. However, even without these enhancements, the current quality of automated subject indexing is sufficient for performing many tasks that previously have relied on laborious human annotation.

Acknowledgements

This work is a part of the national semantic web ontology project FinnONTO[10] 2003-2012, funded mainly by the National Funding Agency for Technology Innovation (Tekes). We thank Hanna Heikkonen and Tuomas Palonen for organizing and coordinating the human indexing of test document sets.

References

1. Buitelaar, P., Declerck, T.: Linguistic Annotation for the Semantic Web. In: Annotation for the Semantic Web, pp. 93–110. IOS Press, Amsterdam, the Netherlands (2003)
2. El-Shishtawy, T., Al-Sammak, A.: Arabic Keyphrase Extraction using Linguistic knowledge and Machine Learning Techniques. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools (2009)
3. Hawking, D., Zobel, J.: Does Topic Metadata Help With Web Search? Journal of the American Society for Information Science and Technology 58(5), 613–628 (2007)
4. Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J.: Unlimited Vocabulary Speech Recognition with Morph Language Models Applied to Finnish. Computer Speech & Language 20(4), 515–541 (2006)
5. Lindén, K., Silfverberg, M., Pirinen, T.: HFST Tools for Morphology – An Efficient Open-Source Package for Construction of Morphological Analyzers. In: Mahlow, C., Piotrowski, M. (eds.) State of the Art in Computational Morphology, Communications in Computer and Information Science, vol. 41, pp. 28–47. Springer Berlin Heidelberg (2009)
6. Löfberg, L., Archer, D., Piao, S., Rayson, P., Mcenery, T., Varantola, K., pekka Juntunen, J.: Porting an English semantic tagger to the Finnish language. In: Proceedings of the Corpus Linguistics 2003 Conference (2003)
7. Löfberg, L., Piao, S., Nykanen, A., Varantola, K., Rayson, P., Juntunen, J.P.: A semantic tagger for the Finnish language. In: Proceedings of Corpus Linguistics 2005 (2005)
8. Markey, K.: Interindexer Consistency Tests: A Literature Review and Report of a Test of Consistency in Indexing Visual Materials. Library and Information Science Research, An International Journal 6(2), 155–77 (1984)

[10] http://www.seco.tkk.fi/projects/finnonto/

9. Maron, M.E.: Automatic Indexing: an Experimental Inquiry. Journal of the ACM (JACM) 8(3), 404–417 (1961)
10. Medelyan, O.: Human-competitive automatic topic indexing. Ph.D. thesis, University of Waikato, Department of Computer Science (2009)
11. Medelyan, O., Witten, I.H.: Thesaurus Based Automatic Keyphrase Indexing. In: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (2006)
12. Oflazer, K., Kuruöz, I.: Tagging and Morphological Disambiguation of Turkish Text. In: Proceedings of the fourth conference on Applied natural language processing (1994)
13. Pala, N., Çiçekli, I.: Turkish Keyphrase Extraction Using KEA. In: Proceedings of the 22nd International Symposium on Computer and Information Sciences (ISCIS 2007) (2007)
14. Pennanen, P., Alatalo, T.: Leiki – a platform for personalized content targeting. In: Proceedings of the 12th ACM conference on Hypertext and Hypermedia (HYPERTEXT'01) (2001)
15. Rolling, L.: Indexing consistency, quality and efficiency. Information Processing & Management 17(2), 69–76 (1981)
16. Saarti, J.: Consistency of subject indexing of novels by public library professionals and patrons. Journal of Documentation 58(1), 49–65 (2002)
17. Salton, G., Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24(5), 513–523 (1988)
18. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 34(1), 1–47 (2002)
19. Tapanainen, P., Järvinen, T.: A non-projective dependency parser. In: Proceedings of the Fifth Conference on Applied Natural Language Processing (1997)
20. Trieschnigg, D., Pezik, P., Lee, V., de Jong, F., Kraaij, W., Rebholz-Schuhmann, D.: MeSH Up: Effective MeSH Text Classification for Improved Document Retrieval. Bioinformatics 25(11), 1412–1418 (2009)
21. Valkeapää, O., Alm, O., Hyvönen, E.: Efficient Content Creation on the Semantic Web Using Metadata Schemas with Domain Ontology Services (System Description). In: Proceedings of the European Semantic Web Conference ESWC 2007, Innsbruck, Austria (2007)
22. Vehviläinen, A., Hyvönen, E., Alm, O.: A semi-automatic semantic annotation and authoring tool for a library help desk service. In: Emerging Technologies for Semantic Work Environments: Techniques, Methods, and Applications, pp. 100–114. IGI Group, Hershey, USA (2008)
23. Witten, I.H., Paynter, G., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA: Practical Automatic Keyphrase Extraction. In: Proceedings of Digital Libraries 99 (1999)
24. Zunde, P., Dexter, M.E.: Indexing Consistency and Quality. American Documentation 20(3), 259–267 (1969)