DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2011

A Language Independent Approach for Aligning Subject Heading Systems with Geographic Ontologies

Nuno Freire[1,2], José Borbinha[1] and Pável Calado[1]
[1]Instituto Superior Técnico, Portugal
[2]The European Library, National Library of the Netherlands, Netherlands
{nuno.freire,jlb,pavel.calado}@ist.utl.pt

Abstract

Subject headings systems are tools for organization of knowledge that have been developed over the years by libraries. The SKOS Simple Knowledge Organization System provides a practical way to represent subject headings systems, and several libraries have taken the initiative to make these systems widely available as open linked data. Each individual subject heading describes a concept, however, in the majority of cases, one subject heading is actually a combination of several concepts, such as a topic bounded in geographical and temporal scopes. In these cases, the label of the concept actually contains several concepts which are not represented in structured form. This paper address the alignment of the geographic concepts described in subject headings systems with their correspondence in geographic ontologies. Our approach first recognizes the place names in the subject headings using entity recognition techniques and follows with the resolution of the place names in a target geographic ontology. The system is based on machine learning and was designed to be language independent so that it can be applied to the many existing subject headings systems. Our approach was evaluated on a subset of the Library of Congress Subject Headings, achieving an F_1 score of 93%.

Keywords: entity recognition; entity resolution; subject headings; linked data; SKOS; machine learning.

1. Introduction

Subject headings systems are tools for organization of knowledge, which have been developed over the years by libraries. Assignment of subject headings to the items within their collections is a part of bibliographic organization tasks carried out by libraries. Subject headings aid the user to discover items in the catalogue that pertain to similar subject matter (Hoerman et al., 2000). Subject headings systems, like other knowledge organization systems such as thesauri and taxonomies, can nowadays be more widely used if made available within the framework of the Semantic Web. The SKOS Simple Knowledge Organization System (Miles et al., 2009) has been developed for this purpose, and it provides a practical way to represent subject headings systems using the Resource Description Framework.

Several libraries have taken the initiative to make subject headings systems widely available by representing them in SKOS and making them available as open linked data. Some known examples are the Library of Congress Subject Headings (LCSH), the *Répertoire d'autorité-matière encyclopédique et alphabétique unifié* (RAMEAU), and *Schlagwortnormdatei* (SWD), which are subject headings systems in English, French and German, respectively.
Each individual subject heading describes a concept. However, in the majority of cases, one subject heading is actually a combination of several concepts, such as a topic bounded in geographical and temporal scopes. Although the concept is available in SKOS, and therefore available with some semantics for machine processing, its individual subtopics are not, which limits what machines can inference from the subject headings.

As subject heading systems become available as open linked data, the value of linking all these sub concepts to their representation in other open data sets becomes more relevant. Several

https://doi.org/10.23106/dcmi.952135661

millions of resources have assigned subject headings, in libraries catalogues and digital libraries. Improving the semantics of subject headings has the potential to benefit the retrieval and access to all these resources. Also, LCSH is one of the Vocabulary Encoding Schemes defined in the DCMI Metadata Terms.

In our work, we address the alignment of the geographic concepts described in subject headings systems with their correspondence in geographic ontologies, such as Geonames[1] and the Getty Thesaurus of Geographic Names[2]. Our approach consists in two tasks, in the first the place names are recognized from the subject headings using a named entity recognition technique developed for the particular case of subject headings. The second task consists in choosing a possible resolution candidate from the target geographic ontology. The system was designed to be language independent so that it can be applied to the many subject headings systems in use throughout the world.

This paper proceeds in Section 2 with a description of the problem. Section 3 summarizes the state of the art in entity recognition and resolution of place names, and Section 4 follows with a description of our approach and details of its implementation. Section 5 presents the evaluation procedure and the obtained results. Section 6 describes the results of the alignment performed on LCSH and RAMEAU. Section 7 concludes and discusses future work.

2. The Problem

Subject headings present a scenario with particular characteristics for the application of information extraction. To make all the concepts within a subject heading available for machine processing with full semantics, their names need first to be recognized through entity recognition techniques. The recognized names then need to be resolved, and disambiguated, if necessary. This section describes how these two steps, and the challenges they present in the particular case of subject headings.

2.1. Recognition of the Places Names

The available entity recognition techniques, when applied to subject headings, are unable to reliably recognize the entities. These techniques are dependent on the lexical evidence provided by well-formed sentences. In subjects heading such lexical evidence is not available, since the headings are a concatenation of simple textual references to concepts. The following are some examples of geographic subject headings from the LCSH:

- Potsdamer Platz (Berlin, Germany)
- Québec (Québec)--History--French and Indian War, 1755-1763
- United States--History--Civil War, 1861-1865--Propaganda
- Cass Lake (Cass County and Beltrami County, Minn. : Lake)
- Portugal--History--Revolution, 1974

In these examples we can observe the heterogeneity of the structure of subject headings. Some delimiting punctuation ("--") is used between the main concepts but they do not provide any clues about the type of the entities that they delimit.

TABLE 1: Examples of entity recognition in subject headings

Québec (Québec)--History--French and Indian War, 1755-1763
[GEO Québec] ([GEO Québec])--[TOPIC History]--[HISTORIC French and Indian War], [TIME 1755-1763]
Portugal--History--Revolution, 1974
[GEO Portugal]—[TOPIC History]--[HISTORIC Revolution], [TIME 1974]
Potsdamer Platz (Berlin, Germany)
[GEO Potsdamer Platz] ([GEO Berlin], [GEO Germany])

[1] http://www.geonames.org/
[2] http://www.getty.edu/research/tools/vocabularies/tgn/index.html

https://doi.org/10.23106/dcmi.952135661

The desired output result of the entity recognition process is the location of the entities and the identification of their type. Table 1 illustrates the desired output, as annotated subject headings for entities of the types: geographical entity, topics, time periods, and historical periods.

In the case of geographical entities, it can be observed that the geographic entity contained in a subject heading may be referred to by its name only, or it may contain additional place names higher in the hierarchy of administrative subdivisions (country, region, state, etc.). It is crucial that the recognition phase is able to identify all these place names, so that the following step of resolution can choose the right candidate from the geographic ontology.

Several approaches can be adapted for this particular scenario of entity recognition. Similar problems have been addressed in many fields, such as bioinformatics, computational linguistics and speech recognition, but the most similar has been the citation matching problem (Wellner et al., 2004) where entity recognition is based on structural characteristics of the text instead of grammatical evidence.

In our work we analysed the available techniques and applied a particular one. The chosen technique is better adapted to capture the structural characteristics of the subject headings, and of the entities in them, in a language independent way.

2.2. Resolution of the Place Names

The resolution of the recognized names consists mainly in two problems: disambiguation of the place name, and estimating the probability that a correct resolution was made.

The ambiguity problems in place names can be characterized according to two types, namely geo/non-geo or geo/geo (Amitay et al., 2004). Geo/non-geo ambiguity refers to the case of place names having other non geographic meanings (e.g., Georgia may refer to the country or to a person). Some common words are, for instance, also place names (e.g., Turkey). On the other hand, geo/geo ambiguity arises when two distinct places have the same name. The geo/non-geo ambiguity is addressed during the recognition phase, while geo/geo ambiguity is addressed while disambiguating the recognized place names.

We have measured the level of geo/geo ambiguity in the place names found in subject headings, by matching the place names found in the geographic subject headings from LCSH and RAMEAU against the place names in the Geonames ontology. We have found 52% of place names to be ambiguous in LCSH, and 71% in RAMEAU. Table 2 shows the detailed results.
In our work we investigated forms of obtaining evidence to support the resolution of the place names, and evaluated techniques for reasoning on the evidence.

TABLE 2: Ambiguity in the place names found in LCSH and RAMEAU (calculated by matching the place names against the places names in Geonames)

	Total Geographic Subject Headings	Ambiguous Place Names				
		2 candidates	3 candidates	4 candidates	5+ candidates	Total
LCSH	61610	6199 (10%)	3381 (5%)	2110 (3%)	20205 (33%)	31895 (52%)
RAMEAU	53301	17283 (32%)	4206 (8%)	2568 (5%)	13738 (26%)	37795 (71%)

3. Related work

Place name recognition concerns the delimiting, in unstructured text, of the character strings that refer to place names. This is a particular instance of the more general problem of Named Entity Recognition (NER), which has been extensively studied in the Natural Language Processing (NLP) community. Place name resolution refers to associating the recognized references into the corresponding entries in a gazetteer (a geographic ontology). This latter sub-task has been addressed by the Geographic Information Retrieval (GIR) community.

The NER task refers to locating and classifying atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of time, quantities,

https://doi.org/10.23106/dcmi.952135661

etc. (Nadeau et al., 2007). Current solutions can achieve near-human performance in well-formed text, achieving F-measure accuracy around 90% (Nadeau et al., 2007).

Initial approaches, which are still commonly used, were based on manually constructed finite state patterns and/or collections of entity names (Nadeau et al., 2007). However, named entity recognition is considered as a typical scenario for the application of machine learning algorithms, because of the potential availability of many types of evidence, which form the input variables (McCallum et al., 2000).

Two particular types of supervised machine learning algorithms have been successfully used for entity recognition. Classification algorithms classify words, or groups of words, according to their entity type. Some examples are Support Vector Machines (Ravin et al., 1997) and Decision Trees (Mikheev, 1999). Nevertheless, the problem was shown to be better solved with sequence labeling algorithms. The earliest sequential classification techniques applied to entity recognition were Hidden Markov Models (Bikel et al., 1997). However this technique does not allow the incorporation into the predictive model of the wide range of evidence that is available for entity recognition. This limitation has led to the application of other algorithms such as the Maximum Entropy Markov Model (McCallum et al., 2000) and Conditional Random Fields (Lafferty et al., 2001). Conditional Random Fields is currently the state of the art for entity recognition.

While NER approaches can be designed to rely entirely on evidence from within the text, place name resolution requires always external knowledge for resolving place names into geographic entities. Geonames is an example of a wide-coverage gazetteer, describing over 7.5 million places from all around the world and has been used in many GIR experiments (Wick et al., 2007).

Similarly to the general case of NER, the main challenges in resolving place names are related to ambiguity. Several approaches for place reference resolution have been proposed in the past. For instance, in Wick et al. (2007), a system to resolve locations mentioned in transcripts of news broadcasts is described. Kanada (1999) reports a value of 96% precision for geographic name disambiguation in Japanese text, with a gazetteer of 96,000 Japanese place names. A variety of approaches have been surveyed in Leidner (2007), where it is concluded that most methods rely on gazetteer matching for performing the identification, together with natural language processing heuristics for performing the disambiguation.

4. The Approach

Our alignment approach consists in a two task process. In the first task, entity recognition is performed on the label of the subject heading, with the purpose of recognizing all references to places. The recognized places are the input for the second task, which will try to resolve them by choosing an entity described in a geographic ontology. This section describes these two tasks.

4.1. Place Name Recognition

We opted for a sequence labelling approach for recognizing entities in subject headings. The core of our approach lies in identifying the most likely sequence of labels for the words and punctuation marks in any given subject heading. The labels used correspond to four entity types (geographical entities, topics, time periods, and historical periods) plus a label for "*not an entity*".

Although our overall system only addresses the geographic entities, we decided to design the entity recognition component to recognize all the entity types that are frequently used in geographic subject headings. The other entity types can potentially be used to support the recognition and resolution of the place names. Also, in future work, these other entities can be addressed and aligned with other ontologies.

In the remainder of this section we will shortly introduce the predictive model for sequence labelling used in our approach, and then describe the specific features for building our model.

https://doi.org/10.23106/dcmi.952135661

4.1.1. The Base Predictive Model

We use as a basis the conditional models of conditional random fields (CRF) which define a conditional probability $p(y|x)$ over label sequences given a particular observation sequence x. These models allow the labelling of an arbitrary sequence x' by choosing the label sequence y' that maximizes the conditional probability $p(y'|x')$. The conditional nature of these models allows arbitrary characteristics of the sequences to be captured by the model, without requiring previous knowledge, by the modeller, about how these characteristics are related.

A CRF is an undirected graphical model that defines a single log-linear distribution over label sequences given a particular observation sequence. The probability of a particular label sequence y given observation sequence x is a normalized product of potential functions, each of the form (Lafferty et al., 2001):

$$\exp(\sum_j \lambda_j t_j (y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k (y_i, x, i))$$

Where $t_j (y_{i-1}, y_i, x, i)$ is a transition feature function of the entire observation sequence and the labels at positions i and $i-1$ in the label sequence; $s_k (y_i, x, i)$ is a state feature function of the label at position i and the observation sequence; and λ_j and μ_k are parameters estimated during supervised training.

When defining feature functions, a set of features $b(x, i)$ is created from the observation sequence. The modeller should choose features which capture those characteristics of the empirical distribution of the training data that should also hold for the model distribution.
Each state feature function uses the value of one of these observation features $b(x, i)$ depending on the current state. Similarly, transition feature functions will use the value of the feature depending on both the previous and current state.

4.1.2. Features for Subject Headings

For our specific problem of recognizing entities in subject headings, the tokens (words and punctuation marks) of the subject heading form our sequence. Based on the tokens, we defined a set of features that express the major characteristic of the representation of the entities, and should allow the construction of a general predictive model.

We opted to use only features that are language independent, so that the predictive model could be applied to subject headings systems in other languages than the one used for building it. For this reason the words in the subject headings are not used themselves as features, as typically is done in natural language processing. Only characteristics of the words, which we considered relatively language independent, are captured by the features. The features are shown in Table 3.

TABLE 3: Features for entity recognition in subject headings

Feature	**Description**
isWord(x, *i*)	1 if the token at position *i* is a word; 0 otherwise
isNumber(x, *i*)	1 if the token at position *i* is a number; 0 otherwise
isCapitalizedWord(x, *i*)	1 if the token at position *i* is a word and the first letter is capitalized; 0 otherwise
isInitial (x, *i*)	1 if the token at position *i* has just one letter; 0 otherwise
isTinyWord(x, *i*)	1 if the token at position *i* is a word with character length == 2; 0 otherwise
isSmallWord(x, *i*)	1 if the token at position *i* is a word with character length >2 and <= 4; 0 otherwise
isYear(x, *i*)	1 if the token at position *i* is a number that maybe a representation of an year; 0 otherwise
headingSection(x, *i*)	Number of previous "--" separators
isWhitespace(x, *i*)	1 if the token at position *i* is a whitespace; 0 otherwise
isHyphen(x, *i*)	1 if the token at position *i* is an hyphen; 0 otherwise

https://doi.org/10.23106/dcmi.952135661

In addition, other features, similar to *isHyphen*(x, *i*), were defined for other punctuation marks: coma, colon, semicolon, period, underscore, open bracket, close bracket, open square bracket, close square bracket, apostrophes and quotation marks.

Additional features are defined in similar way, but they refer to previous or following tokens, instead of the current one. The features: *isWord, isNumber, isWhitespace,* and all *is"PunctuationMark"* are applied also regarding the preceding three tokens, and the following two tokens. In total, the CRF predictive model is based on 96 features.

4.2. Place Name Resolution

The resolution task aims to find a single entity in geographic ontology for the whole subject heading. When more than one place name has been recognized in a subject heading, the resolution system tries to find *part_of* relationships between the place names, since they typically refer to related administrative divisions.

The first step of this task is to find all possible candidates for the resolution, in the geographic ontology, of the first recognized place name. In the next step, features are extracted for each of the resolution candidates. Table 4 presents these features and how they are extracted.

TABLE 4: Feature space for the place name resolution classifier

Feature	Description
Number of words	The number of words in the place name. Place names with more words are more likely to be correctly resolved.
Name match	If the recognized place name matched: the main name of the place; an alternate name in the language of the subject heading; an alternate name without language; an alternate name in another language. If more than one name match, the feature will take the most relevant match.
Exact name match	A boolean value indicating if the recognized place name matched exactly the place name or if not all words matched.
Relative population	A real value between 0 and 1, indicating the relative population of the candidate in comparison with the other candidates. The candidate with the highest population has a value of 1, and 0 corresponds to non-populated places.
Geographic feature type	The type of geographic feature type of the candidate: continent, country, administrative division, island, city, natural landmark (rivers, mountains, forests, etc.), and *other*. We opted to use this simpler list to allow independence from the geographic ontology.
Related places found	The number of other place names recognized in the subject heading (typically administrative divisions) which were found in the administrative hierarchy of the candidate.
Relative related places	A real value between 0 and 1, indicating the relative number of administrative divisions found in the subject heading, in comparison to the other candidates.
In source country	A boolean value stating if the candidate is located in one of the source countries of the subject heading system. In the subject headings, the country of the place is often omitted when it is located in the country where the subject heading system is maintained. For LCSH, we used as source countries: The United States of America, Canada and Australia. For RAMEAU, we used France.

For choosing a resolution candidate for the subject heading, reasoning is performed on the extracted features, and each candidate is classified as *match* or *non-match*. The classifier provides the probability of each of the candidates being the correct one, and the one with the highest probability is chosen. If none of the resolution candidates achieves a minimum probability threshold for the class *match*, no alignment is established.

The resolution reasoning is implemented through a machine learned classifier component. Several machine leaning classification algorithms were evaluated and compared, including Support Vector Machines, Decision Trees, Random Forests and Bayesian networks. We have chosen to use a Decision Tree induced using the C4.5 (Quinlan, 1993) algorithm, since it consistently resulted in higher results in the F_1-measure and lower mean absolute error, in cross-validation tests. The C4.5 algorithm was configured for building a Decision Tree with a maximum depth of 15, and perform pruning on the final tree.

https://doi.org/10.23106/dcmi.952135661

5. Evaluation

An evaluation of the alignment approach was performed on a subset of the Library of Congress Subject Headings. A random selection of 800 subject headings was made from subject headings whose main concept was geographic. This data set was used to evaluate the two subtasks of our approach independently. It was also used to evaluate the final alignment results.
For the evaluation of the entity recognition task, the subject headings were manually annotated. All entities in the label of the concept were identified and annotated with the corresponding type: geographical entities, topics, time periods, and historical periods. Table 5 summarizes the amount of entities found for each entity type. In total, 1985 entities were found in the 800 subject headings, resulting on an average of 2.48 entities per subject heading.

TABLE 5: Annotated entities in the LCSH evaluation data set

Subject Headings	Geographical Entities	Topics	Time	Historical Periods	Total Entities
800	1348	371	200	66	1985

For the evaluation of the resolution task, the subject headings were manually aligned with Geonames. For the 800 subject headings, 262 (33%) headings had no correspondence in Geonames, therefore they were not used for the evaluation of the resolution task.

This section follows with the presentation of the results of both tasks and the final alignment. These where evaluated according to the measured *precision* (the percentage of correct results in all results found), *recall* (the percentage of entities found compared to all existing entities), and *F_1-measure* (the weighted harmonic mean of precision and recall).

5.1. Place Name Recognition

For the evaluation method we have chosen used the *exact-match* method, which has been used in several named entity recognition evaluation tasks, such as the Conference on Natural Language Learning (Sang et al., 2003). In the *exact-match* method, an entity is only considered correctly recognized when it is exactly located as in the manual annotation. Recognition of only part of the name, or with words that are not part of the name, is not considered correct.

The evaluation was performed as a cross-validation test, which involves partitioning the evaluation data set into complementary subsets of the data set, testing the classifier on one subset, while training it on the remaining subset. 10-fold cross-validation was performed, and the validation results were averaged over the ten runs. The results obtained, broken down by entity type, are show in Table 6.

TABLE 6: Entity recognition precision, recall and F_1-measure, measured using 10-fold cross-validation

Entity Type	Precision	Recall	F_1-measure
Geographical entities	0.981	0.978	0.980
Topics	0.981	0.970	0.976
Time	0.985	0.985	0.985
Historical Periods	0.942	0.985	0.963
All Entities	**0.980**	**0.978**	**0.979**

We consider the results obtained to be good indication that entities can be reliably recognized in subject headings in a language independent way, and that the CRF based predictive model was able to capture the patterns in the data, achieving an overall F_1-measure of 0.979.

5.2. Place Name Resolution

The resolution of the place names in Geonames was evaluated by two methods. The first method aimed to evaluate the contribution of each individual feature for the resolution process.

https://doi.org/10.23106/dcmi.952135661

The second method aimed to evaluate the overall quality of the machine learned classification model.

Evaluation of the individual features was performed using the wrapper methodology, (Kohavi et al., 1997). This method consists in using the prediction performance of the learning machine to assess the relative usefulness of subsets of features.

An exhaustive search of all feature combinations was performed. Each combination was evaluated by a 10-fold cross-validation test and the best performing feature combination was noted. Table 7 summarizes the results, by showing the features that were present in the best performing combination of the 10 folds. Since all features contributed to the correct resolution in at least one fold, we used them all in the final system.

TABLE 7: Evaluation results of the features for resolution

Feature	Number of folds (%)
Relative population	100%
Related places found	90%
Name match	80%
Number of words	80%
Relative related places	70%
Exact name match	40%
Geographic feature type	30%
In source country	20%

The overall quality of the machine learned classification model was evaluated by measuring precision, recall and F_1-measure on a 10-fold cross-validation test on the evaluation data set.

In this evaluation, each resolution candidate was tested independently of all others, unlike in the final system, where only the highest probability candidate is chosen. A candidate was considered valid if the probability given by the resolution classifier for the class *match* is higher than a minimum threshold (the minimum confidence level). The measurements were taken at six levels of minimum confidence, and they are shown in Figure 1.

We consider the results to be a positive indication that the predictive model is not over fitting the training data. The measured value for the Mean Absolute Error (0.016) also supports this analysis.

The system is able to achieve very high precision of 97% at 40% recall, or reach a recall of 83% at 85% precision. Summing up, we conclude the following from the cross-validation evaluation:

- The classifier is not over fitting to the training data;
- A very high precision is possible to obtain, but at the expense of recall;
- Recall never reaches very high levels, and lowers considerably if high precision is required

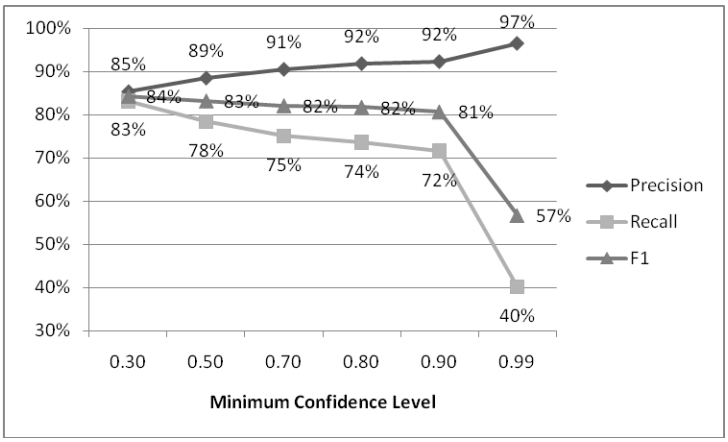

FIG. 1. Place name resolution precision, recall and F_1-measure, measured using 10-fold cross-validation

https://doi.org/10.23106/dcmi.952135661

5.3. Alignment Evaluation

The complete approach, including the two tasks of recognition and resolution, was evaluated by a 10-fold cross-validation test on the evaluation data set.

An alignment from a subject heading to a Geonames entity is only established if the resolution candidate with highest probability, given by the resolution classifier, is higher than a chosen minimum confidence level. The measurements were taken at six levels of minimum confidence, and they are shown in Figure 2. Note that for the calculation of the recall, only those subject headings that had corresponding entry in Geonames were considered.

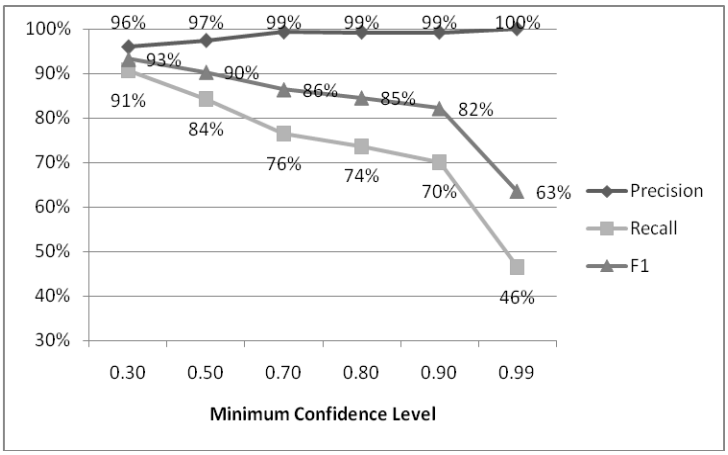

FIG. 2. Alignment precision, recall and F_1-measure, measured using 10-fold cross-validation

The system is able to achieve very high precision of 99% with 76% recall. Perfect precision was achieved, but with a steep loss of recall to 46%. The F_1-measure shows that small increases in precision, obtained by increasing the minimum confidence threshold, lead to higher relative losses in recall. The best F_1-measure result of 93% was obtained on the lowest confidence level. Summing up, we conclude that very high precision is possible, leading to very reliable alignments, and that recall lowers considerably if high precision is required. Applications should therefore choose the appropriate confidence level for their own objectives.

6. Alignment Results

In this section we present the results obtained by applying the system on two subject heading systems in different languages: LCSH, in English, and RAMEAU, in French. Both systems were processed in their SKOS representations.

All geographic subject headings of both systems were processed for alignment with Geonames. The number of alignments established, at different minimum confidence levels, is shown in Table 8. The corresponding percentages of alignments are shown in Figure 3.

The percentage of alignments was slightly lower than expected. On the evaluation data set we observed that 33% of LCSH geographic subject headings had no alignment in Geonames, and the measured recall ranged from 46% to 91%. Therefore, we expected to find alignments for 30% to 60% of the geographic subject headings, but found 18-56% in RAMEAU, and 24-57% in LCSH. This difference however, is not statistically significant (P>0.10).

https://doi.org/10.23106/dcmi.952135661

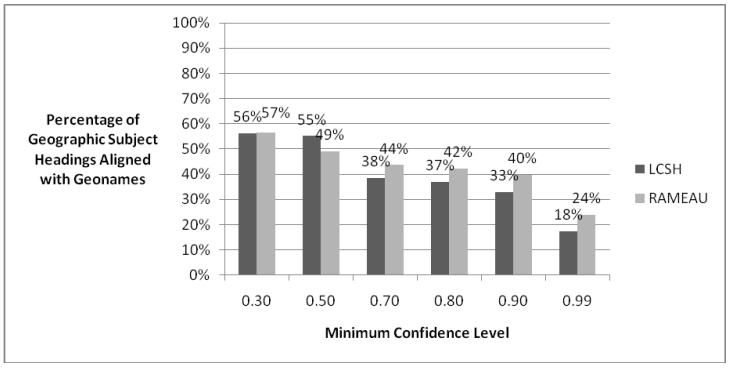

FIG. 3. Percentage of geographic subject headings from LCSH and RAMEAU aligned with Geonames

TABLE 8: Total geographic subject headings from LCSH and RAMEAU aligned with Geonames

	Geographic Subject Headings	Subject Headings Aligned with Geonames					
		confidence >= 0.3	confidence >= 0.5	confidence >= 0.7	confidence >= 0.8	confidence >= 0.9	confidence >= 0.99
LCSH	61610	34912	30281	26885	25976	24518	14737
RAMEAU	53301	30006	29412	20517	19726	17492	9329

7. Conclusions and Future Work

This paper presented an approach for aligning the geographic concepts described in subject headings systems with their correspondence in geographic ontologies. The approach was designed to be language independent, having in mind its general applicability to subject heading systems in any language. Our approach achieved a maximum F_1-measure of 93%.

This work is a first step towards improving the semantics of the concepts represented in subject heading and available in SKOS as open linked data. The positive results of our approach provide a foundation towards establishing alignments for other entity types besides geographic.

We also expect that, the outcome of this work can be used to automatically establish links between subject headings systems in different languages, such as the work carried out in the MACS project[3] and related research (Isaac et al., 2008).

References

Amitay, E., Har'El, N., Sivan, R., Soffer, A. (2004). Web-a-where: geotagging web content. In Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval.

Bikel, D., Daniel, M., Miller, S., Schwartz, R., Weischedel, R. (1997). Nymble: a High-Performance Learning Name-finder. Proceedings of the Conference on Applied Natural Language Processing.

Hoerman, H.L., Furniss, K. A. (2000). Turning Practice into Principles: A Comparison of the IFLA Principles Underlying Subject Heading Languages (SHLs) and the Principles Underlying the Library of Congress Subject Headings System. The Haworth Press, Inc., Cataloging & Classification Quarterly, vol. 29, no. 1/2, 31-52.

Isaac, A., Matthezing, H., Schlobach, S., Zinn, C. (2008). Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies. Library Review, 57.

Kanada, Y. (1999). A method of geographical name extraction from Japanese text for thematic geographical search. In proceedings of the 8th International Conference on Information and Knowledge Management.

Kohavi , R., G. John. (1997). Wrappers for feature selection. Artificial Intelligence, 97(1-2):273–324.

Lafferty, J., McCallum, A., Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In International Conference on Machine Learning.

Leidner, J. (2007). Toponym Resolution in Text. PhD thesis, University of Edinburgh.

McCallum, A., Freitag, D., Pereira, F.. (2000). Maximum entropy Markov models for information extraction and segmentation. International Conference on Machine Learning

[3] https://macs.hoppie.nl/pub/

https://doi.org/10.23106/dcmi.952135661

Mikheev, A. (1999). A Knowledge-free Method for Capitalized Word Disambiguation. In the 37th annual meeting of the association for computational linguistics, 159-166.

Miles, A.J., Bechhofer, S. (2009). SKOS Reference. W3C Recommendation. Latest version available at http://www.w3.org/TR/skos- reference/..

Nadeau, D., S. Sekine. (2007). A survey of named entity recognition and classification. Linguisticae Investigationes 30.

Quinlan, J.R. (1993). C4.5: Programs for Machine Learning, Morgan Kauffman.

Ravin, Y., Wacholder, N. (1997). Extracting Names from Natural-Language Text.

Sang, T.K., F. Erik, F. De Meulder. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings Conference on Natural Language Learning.

Wellner, B., McCallum, A., Peng, F., Hay, M.. (2004). An Integrated, Conditional Model of Information Extraction and Coreference with Application to Citation Matching. UAI '04 Proceedings of the 20th conference on Uncertainty in artificial intelligence.

Wick, M., T. Becker. (2007). Enhancing RSS Feeds with Extracted Geospatial Information for Further Processing and Visualization. In The Geospatial Web - How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society, Springer.

https://doi.org/10.23106/dcmi.952135661