Multimed Tools Appl (2014) 68:413–427
DOI 10.1007/s11042-012-1339-y

Annotating korean text documents with linked data resources

David Müller · Mun Yong Yi

Published online: 30 January 2013

Abstract Semantic annotation approaches link entities from a knowledge base to mentions of entities in text to provide additional content-related information. Recently increasing use of resources from the Linked Open Data (LOD) Cloud has been made to annotate text documents thanks to the network of machine-understandable, interlinked data. While existing approaches to semantic annotation in the LOD context have been proven to be well performing with the English language, many other languages in general and the Korean language in particular are still underrepresented. We investigate the applicability of existing semantic annotation approaches to the Korean language by adapting two popular approaches in the semantic annotation field and evaluating those approaches on an English-Korean bilingual sense-tagged corpus. Further, general challenges in internationalization of annotation approaches are summarized.

Keywords Semantic annotation · Entity linking · Linked data · Korean · LOD

1 Introduction

Dealing with multilingual resources is considered a major challenge for the Semantic Web [3]. The internationalization and localization of Semantic Web applications has recently become the focus of intensive research with the Linked Open Data 2[1] project promising to make further contributions towards the interlinking and fusion

[1] http://lod2.eu

D. Müller
Karlsruhe Institute of Technology, Karlsruhe, Germany
e-mail: mueller@nate.com

M. Y. Yi (✉)
Department of Knowledge Service Engineering, KAIST, Daejeon, Korea
e-mail: munyi@kaist.ac.kr

of multilingual resources. In that context, the integration of Asian languages and resources into existing technologies in the field poses several special difficulties, which have not yet all been sufficiently addressed [2].

The Linked Open Data (LOD) project as a central initiative of the Semantic Web community established best practices for connecting structured data on the Web, which lead to a steadily growing number of interlinked datasets - referred to as the Linked Open Data Cloud[2] [11].

Semantic annotation refers to linking entities from a knowledge base to mentions of entities in text in order to provide additional content-related information (names, attributes and descriptions). Semantic annotation systems have recently made increasing use of resources from the Linked Open Data Cloud to annotate text documents benefiting from a network of machine-understandable, interlinked data and shared URIs [15].

In a multimedia context, semantic annotation approaches promise to be of benefit in the automatic monitoring of online social media streams, real-time event detection, the provision of context-related metadata or targeted advertisements through the identification of topics in unstructured text resources on web pages.

DBpedia[3] as an approach to turn the content of Wikipedia into structured knowledge using Semantic Web technologies has been established as the central hub of the Linked Open Data Cloud. Wikipedia content is extracted, converted to RDF and interlinked with other LOD resources. DBpedia provides several interfaces for access to the structured Wikipedia data [1]. The central position of DBpedia within the LOD Cloud and its domain independency make it a well suited starting point for semantic annotation in the LOD context.

Semantic annotation with DBpedia resources [1] and the discovery of related information specified in the LOD Cloud through DBpedia data is exemplified in Fig. 1.[4] Linking *President Obama* in text to the corresponding DBpedia resource `Presidency_of_Barack_Obama` enables discovering additional content-related information through the links specified in the DBpedia data. In the given example, we identify *President Obama* to reside in the `White_House` and further locate the `White_House` at a set of given geological coordinates. By following the links and consuming the specified data we can now not only assume that the specified text is related to the given geological coordinates, we also understand the nature of this relation.

In state of the art semantic annotation approaches in the LOD context an obvious problem comes to surface when reviewing systems regarding their language support: While semantic annotation approaches with English language support are various, current approaches in semantic annotation are still limited to a small set of supported input languages and evaluation efforts so far were only targeting English language performance. Furthermore, the annotation of Korean language text with LOD resources has not been the subject of academic research up to this point.

Several reasons may exist why the problems mentioned above have not been solved before. Annotation methods rely on a set of language-specific input data, and

[2] http://thedatahub.org/group/lodcloud

[3] http://dbpedia.org

[4] Example sentence taken from washingtonpost.com

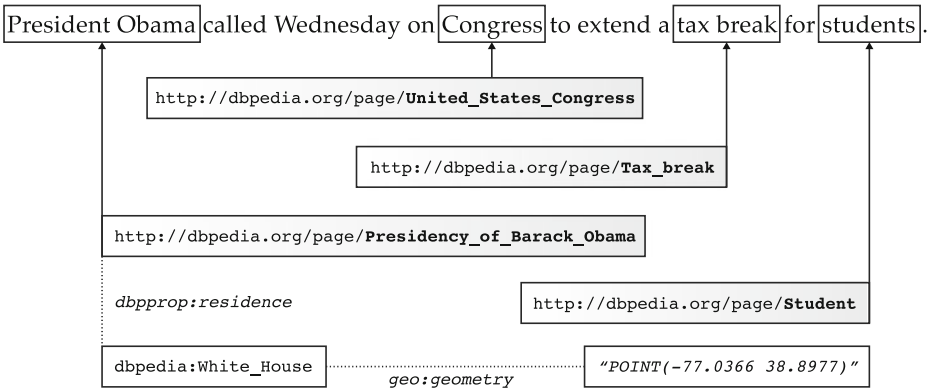

Fig. 1 Introductory example to semantic annotation

processing technology and algorithms need to be optimized for the relevant language to produce best results. Existing research suggests that the localization of semantic annotation systems that were optimized for English language to non-latin languages is not trivial, since semantic annotation efforts depend on the progress research has made in related fields such as language-related morphological and syntactic processing technology and the language-specific availability of knowledge bases in general and language-specific data sets linked to the LOD Cloud in particular [2, 4, 6]. Research in the field does in consequence depend on previous work in several areas which have achieved higher attention in the English language research community for reasons being of economic and historic nature [4]. We however believe that with the recent efforts to solve issues of multilingualism in the Semantic Web community and the introduction of the Korean DBpedia [12], it is time to review the field of semantic annotation from a specific language perspective.

This study investigates the applicability of existing semantic annotation approaches in the LOD context to the Korean language by adapting and evaluating popular approaches in the field and summarizing the general challenges in internationalization and localization of annotation approaches.

2 Related work

The remainder of this work requires knowledge of the Semantic Web in general and concepts and ideas behind Linked Data in particular. A thorough summary of Linked Data and related technologies has been given by Heath et al. [11].

While the topic of semantic annotation in general has been the subject of a broad range of academic research and a great number of systems have emerged, most systems have been limited to domain-specific annotation vocabularies that do not qualify for application in the LOD context [9, 15]. The focus of this work shall further be on systems that have no such limitations.

Recently, annotation systems that use text from Wikipedia have emerged as state of the art in semantic annotation [15]. The annotation of text documents with

Wikipedia articles has first been introduced by Mihalcea et. al. in 2007 in their Wikify approach [16]. The Wikify system uses a two-stage approach to annotation, in which the first detection stage makes use of the probability an article is linked if mentioned within Wikipedia and the next disambiguation phase ensures that links are made to the appropriate article based on words and phrases surrounding article links. In 2008, Medelyan et al. [13] further enhanced the Wikify approach by adding an additional stage that identified most important topics by filtering lesser important article links. The disambiguation stage is further simplified by introducing the concepts of commonness (prior probability) and relatedness (relation to other mentioned concepts). Milne and Witten [17] improved the approach proposed by Medelyan by applying machine learning techniques using commonness and relatedness as features as well as making use of contextual information. In 2011, Ratinov et al. [19] proposed to improve Wikification systems by not only disambiguating entities on a global level where all mentions of the same entities detected in a document are grouped leading to the decision to either annotate all mentions of a specific entity or none, but also separately disambiguating each mention locally.

Recently approaches that focus on the annotation of shorter text fragments with Wikipedia articles have surfaced [8, 10]. Ferragina et al. proposed TAGME, which uses a voting scheme for disambiguation where each anchor link votes for candidate annotations of surrounding anchors in text [8]. Meij et al. evaluated current approaches on short text fragments and proposed their own algorithm applying a set of state-of-the-art machine learning techniques [10].

While much research has been directed towards semantic annotation of text documents with Wikipedia articles, fewer researchers have proposed systems that natively annotate text with Linked Data entities [9]. To our best knowledge, DBpedia Spotlight,[5] introduced by Mendes et al. [15] in 2011, is the only system available under open source license that annotates text with Linked Data. Mendes et al. proposed a four step approach to annotation of text documents with DBpedia URIs. In the first spotting stage, phrases in text are detected that may be linked to a DBpedia resource. The second candidate selection stage finds a set of candidate resources in DBpedia that link to detected phrases, followed by the third disambiguation stage that uses Term Frequency and Inverse Candidate Frequency weights (TCF*IF). In the final stage topics with low relevancy are filtered.

Further systems exist that have neither been published in academic context, nor are available under open source license. An overview of state of the art annotation services with high visibility in related work, public availability and applicability to the LOD context is provided in Table 1.

Recent work has evaluated the performance of semantic annotation systems on English language text [8, 10, 15, 20]. Current approaches in the semantic annotation field are limited to a small set of input languages. While no semantic annotation system with Korean language support exists in the LOD context, the DBpedia Spotlight approach [15] and Milne and Witten's approach [17] offer the possibility for adaption to a large set of languages, as both approaches are published under open source license with input data readily available.

[5] http://dbpedia.org/spotlight

Table 1 Publicly available annotation services with LOD support

Name	URL	Reference
AlchemyAPI	www.alchemyapi.com	
DBpedia Spotlight	dbpedia.org/spotlight	[15]
M&W Wikifier	wdm.cs.waikato.ac.nz	[17]
Ontos	www.ontos.com	
OpenCalais	www.opencalais.com	
TAGME	tagme.di.unipi.it	[8]
The Wiki Machine	thewikimachine.fbk.eu	
Yahoo Content Analysis	developer.yahoo.com/contentanalysis	
Zemanta	www.zemanta.com	

Only a few approaches exist for semantic annotation of Korean text [4, 5, 21]. Chai et al. [4, 5] present an annotation system for Korean language based on the EXCOM approach [7] annotating text with semantic categories. Zheng et al. present an approach that can be trained to identify mentions of restaurants in Korean text.

The adaption of semantic annotation approaches to the Korean language poses difficulties that are in part covered by related work [4, 6]. Chai [4] mentions flexible sentence patterns, a complex affix system, and conjugational endings on verbs and adjectives to be major difficulties in Korean language processing. Chung et al. [6] refer to the unconstrained foundation of compound nouns, the large number of possible verb endings, and "long-distance scrambling" as challenges in syntactic parsing.

3 General architecture of semantic annotation systems

For a comparative analysis of semantic annotation systems and language-specific challenges it is necessary to identify a common system structure, which to our best knowledge has not been attempted before. Figure 2 provides an overview of the system architecture shared by all approaches in the field.

We identified semantic annotation systems to consist of 3 major components: 1. indexing component, 2. annotation data source, 3. annotation component. The indexing component extracts data relevant to the annotation process stored in a knowledge base. This data is saved in a data source for quick access by the third annotation component. The annotation component is responsible for the actual annotation process receiving plain text as input and annotating the input text with entities from the knowledge base. This is done in four consecutive steps: (1) text preprocessing, (2) candidate selection, (3) entity disambiguation, and (4) entity annotation. Each of the previously mentioned steps consists of several sub-steps. In text preprocessing, the input text is split into smaller units (tokens) for processing (tokenization), which optionally are assigned to a grammatical category (pos tagging) and reduced to their basic forms (lemmatization). In the next step referred to as candidate selection, tokens and combinations of tokens are matched with entities in the knowledge base, which may represent the meaning of each token or a combination of tokens in text through lookups to the annotation datasource (entity detection), and retrieved candidate entities with a low preliminary probability to be a correct annotation are filtered (filtering). The entity disambiguation step determines

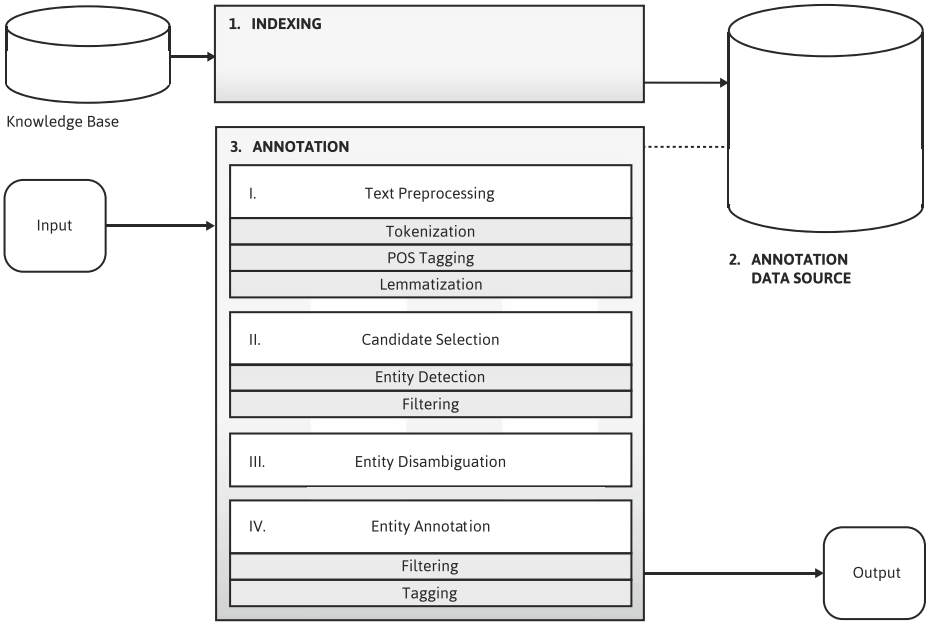

Fig. 2 General architecture of semantic annotation systems

the correct entity matching each token or set of tokens among the set of candidate entities found in the previous step. Finally the fourth entity annotation step filters annotations along various system specific criteria and outputs the annotated text.

4 Language-specific challenges to semantic annotation

Our review of state-of-the-art semantic annotation approaches has revealed that current systems natively support only a small set of languages. Reasons for this lack in language support are attributed to language-specific challenges in system adaption. We identified challenges in two specific categories: (1) *Language processing* referring to the general availability and quality of language-related processing technology and restricted applicability of existing approaches to new language environments, and (2) *knowledge base*, the language-specific availability of knowledge base and quality of available knowledge bases referring to their size, semantic richness, links and lexicalizations (string representations of entities). An overview of identified language-specific challenges and their occurrences in the annotation process is given by Fig. 3. The implications of Fig. 3 are further discussed in the following paragraphs.

4.1 Language processing

Existing approaches make wide use of language processing technology which has been optimized for usage with a small set of languages. Two language processing

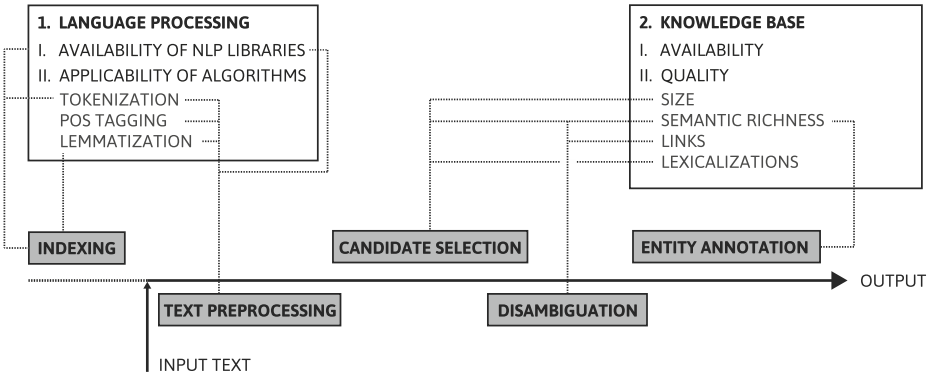

Fig. 3 Language-specific challenges in the annotation process

steps with high visibility in current approaches are *tokenization* (splitting up the input text into several smaller units for processing) and *part-of-speech tagging* (assigning grammatical categories to words). Language processing is needed in both the indexing of knowledge sources and input text preprocessing during annotation. Tokenization and POS tagging require language-specific approaches for optimal performance.

In the selection of candidate annotations for concepts in text, words which fail to be correctly tokenized in the prior text preprocessing step might not be linked to related candidates in the knowledge base. Tokenization functionality in existing

Table 2 Example: Mismatch through whitespace tokenization of Korean text
버락 오바마 미국 대통령이 서울 핵안보정상회의 참석차 한국에 도착하는 첫날 최전방 비무 장지대를 방문한다.
Barack Obama President of the USA visits the front-line Demilitarized Zone the first day he arrives in Korea to attend the Seoul Nuclear Summit.

Token	Retrievable	Not retrievable	Suffix	Translation
버락	버락			Barack
오바마	오바마			Obama
미국	미국			USA
대통령이		대통령	이	President + *subject marker*
서울	서울			Seoul
핵안보정상회의	핵안보정상회의			Nuclear summit
참석차		참석	차	Attendance + *intention marker*
한국에		한국	에	Korea + *direction marker*
도착하는		도착	하는	arrival + *to do*
첫날				first day
최전방				front-line
비무장지대를		비무장지대	를	Demilitarized zone + *object marker*
방문한다		방문	한다	visit + *to do*

Table 3 Comparison of the English and Korean Wikipedia and DBpedia

	English Wikipedia	Korean Wikipedia
Content pages	3,854,657	187,186
Average edits per page	19.66	15.54
Active registered users	132,730	2049
Active registered users per page	0.0344	0.0109
Fraction of articles ≥ 0.5 kb readable text	0.90	0.60
Fraction of articles ≥ 2.0 kb readable text	0.45	0.16
Number of internal links	78.3M	2.4M
Avg. number of internal links per page	25.258	19.67
	English DBpedia	Korean DBpedia
Number of entities	4,236,434	51,566
Number of triples	1,200,000,000	5,451,860
Avg. number of triples per entity	283.28	105.73

approaches [15, 17] primarily based on the splitting of sentences into tokens by whitespaces is not applicable to languages making wide use of suffixes or prefixes on words such as the Korean language [4]. This is shown in the example in Table 2 from recent news media.[6]

In semantic annotation, part-of-speech-tagging (POS tagging) is applied to filter candidate words or phrases in text for annotation by their grammatical category. Algorithms and assigned grammatical categories in POS tagging are highly language-specific. The POS functionality in widely used NLP libraries however is often limited to a small set of languages. Research in the field for many languages including Korean still lags behind efforts for the English language [4].

4.2 Knowledge source

The general availability and specific quality of knowledge sources are of central importance to annotation approaches. Small knowledge sources provide less candidate entities for linking (*size*). The recent focus of semantic annotation approaches on Wikipedia data restricts applicability to language communities where Wikipedia has achieved higher popularity. Features used to determine the general importance of entities inside a network of interlinked entities are dependent on the *semantic richness* of information incorporated in the knowledge source. Outgoing and incoming *links* and the quality of textual content attached to entities influence the level of which relatedness between two entities can be determined. Encyclopedias are further a valuable source of *lexicalizations*. The access to a quality source of lexicalization data is considered a major challenge in developing language processing tools in general and approaches for the Korean language in particular [14].

The English and Korean language versions of Wikipedia and DBpedia show significant differences in size and incorporated information. Table 3 compares the English and Korean language versions of Wikipedia and DBpedia.[7] The fraction of

[6]Source: www.etoday.co.kr (March 22, 2012)

[7]http://ko.dbpedia.org

Wikipedia articles with more than 0.5 kilobytes and 2 kilobytes of text is in each case significantly lower for the Korean language Wikipedia, despite its smaller size (data from January 2010). A comparison of our sense annotated bilingual corpus data has furthermore shown that each kilobyte of English language text incorporates more meaningful text than each kilobyte of its Korean counterpart. The average of internal links per page in the Korean language version is lower than that of the English Wikipedia, a fact that is further confirmed by comparing the average number of triples per entity in the DBpedia language versions.

5 System adaption

Two annotation systems with the possibility for Korean language adaption were previously identified. The approaches by Mendes et al. [15] and Milne and Witten [17] were adapted for Korean language use to examine the applicability of existing solutions in the field of semantic annotation to Korean language, following the documentation for localization efforts published by the authors. Language-specific decisions made in adaption are briefly summarized in the following paragraphs.

5.1 DBpedia Spotlight

The DBpedia Spotlight approach [15] relies on input data from Wikipedia and DBpedia. A Korean language version of DBpedia has recently been introduced.[8] Spotlight uses Apache Lucene[9] to index both datasets. Lucene in its native version does not supply appropriate Korean language functionality. However, an extension to Lucene for the Korean language[10] is available and was integrated into our Korean Spotlight approach. The detection of candidate entities in the input text in Spotlight is performed using LingPipe's dictionary-based chunking approach[11] making use of regular-expression based tokenization and Hidden Markov Model model-based POS tagging to optionally limit candidates to a set of grammatical categories. We replaced the tokenization approach within LingPipe, which is not applicable to Korean language processing with the previously introduced Lucene Korean language extension, and refrained from using LingPipe's POS tagging functionality, which is not easily compatible with Korean language input, considering the latter to rather be of importance in processing time reduction.

5.2 M&W Wikifier

Milne and Witten's approach to semantic annotation is published as part of the WikipediaMiner toolkit [18], which requires a language-specific Wikipedia dump[12] and further language-specific information on the used Wikipedia version as input.

[8] http://ko.dbpedia.org

[9] http://lucene.apache.org

[10] http://sourceforge.net/projects/lucenekorean

[11] http://alias-i.com/lingpipe/docs/api/com/aliasi/dict/ExactDictionaryChunker.html

[12] http://dumps.wikimedia.org/kowiki

The toolkit's indexing functionality makes use of an Apache OpenNLP[13] model for sentence detection currently not available for the Korean language. We trained a new OpenNLP model for Korean language sentence detection based on a corpus containing 100,000 formatted sentences from Korean news articles available over the Leipzig Corpora Collection.[14] The approach uses the WEKA workbench[15] to train classifiers for disambiguation and entity annotation. Classifiers were trained based on a subset of articles from the Korean Wikipedia.

6 Evaluation and discussion

Adapted semantic annotation approaches were evaluated and compared to the performance of their respective English versions on a English-Korean bilingual corpus, which was manually annotated with Wikipedia articles as gold standard. The following paragraphs further describe our experiment setup.

6.1 Sense-tagged corpus

A bilingual sense-tagged corpus for English-Korean is to our best knowledge not available at this point. We chose to manually annotate transcripts from top ranked presentations held at TED conferences,[16] which are freely available for a large set of languages. This approach offers the possibility to extend the evaluation to further languages at a later point in time. The transcripts cover a broad range of topics. We acquired and manually annotated the first three paragraphs of the top ten presentations with available transcripts, following modified guidelines for annotation of Wikipedia articles[17] with repeated annotation of detected articles.

6.2 Measures

Measures for the performance of unranked information retrieval have evolved as the standard in evaluation of semantic annotation systems [8, 10, 15, 20]. For comparability with existing results, we chose to evaluate the performance of annotation systems with *precision*, *recall*, and *balanced F-Measure* (F_1-Score) evaluating each system's annotation output on the previously introduced gold standard. Furthermore, the decision was made to rank results by system specific confidence values and evaluate the performance using *precision at 5* (P_5) and *precision at 10* (P_{10}) measures, accounting for the fact that, without a restriction of annotation output based on confidence values, DBpedia Spotlight tends to create a higher number of annotations ($68.22 \gg 37.22$), leading to a potentially higher recall and lower precision values.

[13] http://opennlp.apache.org

[14] http://corpora.uni-leipzig.de

[15] http://www.cs.waikato.ac.nz/ml/weka

[16] http://www.ted.com

[17] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking

Table 4 Evaluation results

	Milne&Witten [en]	Milne&Witten [ko]	Spotlight [en]	Spotlight [ko]
E	44.20	50.70	44.20	50.70
E_{ret}	37.22	19.56	68.22	31.00
E_{rel}	24.56	06.33	26.11	20.56
Precision	0.683	0.324	0.380	0.663
Recall	0.554	0.121	0.588	0.405
F_1-Score	0.608	0.172	0.460	0.497
P_5	0.900	0.500	0.400	0.867
P_{10}	0.810	0.440	0.433	0.800

6.3 Results

The results of our evaluation assessing the performance of Milne and Witten's Wikifier and the DBpedia Spotlight approach for the English and Korean language in form of the arithmetic average of entities manually annotated in the gold standard E, the system's retrieved entities E_{ret} and retrieved relevant entities E_{rel}, the weighted average values for precision, recall and F1-score over our set of input documents weighted by the total number of entity annotations E_i in the respective gold standard text file i, and the arithmetic average of precision at 5 and precision at 10 values are summarized in Table 4.

6.4 Discussion

The results of our evaluation approach presented above provide a good insight into system-specific performance and to some extent reflect our expectations, while still incorporating surprises and allowing useful insights, which are discussed in the next sections.

The overall best performance of evaluated systems was shown by Milne & Witten's approach for the English language outperforming other systems on all performance measures except recall. To our surprise, a close second best was the Korean DBpedia Spotlight approach, which showed a better performance than its English language counterpart. The English DBpedia Spotlight approach however achieved the highest recall in the field of evaluated systems, which was partly caused by the highest number of average annotations (68.22). Milne & Witten's approach adapted for Korean language use showed the worst performance.

In general, a system's performance is highly influenced by the number of set annotations. A high number of annotation leads to a higher retrieval and lower precision. The average number of annotations made by each system was highly varying on a system and language level. To eliminate the influence of shown tendencies towards either a high recall or a high precision, precision at rank 5 and 10 was used to only assess a system's top 5 and 10 annotations. The DBpedia Spotlight approach for the English language despite retrieving the highest number of annotations performed worst in this measure showing a high number of wrong annotations with high confidence values. Taking a closer look at annotated topics, we have found that the English DBpedia Spotlight tends to wrongly link proper names to rather general concepts with high confidence. This is not visible in the Korean language version of Spotlight, which may partly be due to the fact that in the small

knowledge source in the approach (Korean DBpedia) most of those topics are not present. Other approaches assessing only top ranked results performed as expected leading to precision values above the system's average precision.

Comparing the English language results with existing evaluation approaches of English language annotation systems, the results seem to confirm the tendency towards a slightly better performance of Milne & Witten's approach found by Meij et al. [10].

The importance of appropriate language processing technology on system performance becomes obvious in the performance differences of Milne & Witten's approach, which makes use of a tokenization algorithm that does not perfectly work with Korean language input. The weak performance might partly be attributed to the supervised learning approaches applied in [17], which learns disambiguation and entity annotation tasks based on human annotation in articles that are believed to be less appropriate in the Korean Wikipedia. The raw amount of links per Wikipedia article has been proven to be less of a language-specific factor with the Korean Wikipedia having on average 19.67 outgoing links per article opposed to 25.26 average links in the English language version.

Table 5 shows the estimated fraction of entities in the gold standard that are retrieved with the tokenization approach applied by Milne & Witten in both the English and Korean text corpus (annotated words without prefix or suffix). While for the English language all entities in text are detected, a fraction of only slightly less than 20 % of all entities are correctly tokenized in the Korean text. It becomes obvious that the language processing technology that is applied in data indexing and input text processing has a strong influence on output results. Mistokenization of entities leads to wrong data in indexing and causes failure of annotation of respective entities in input text.

Refraining from POS tagging in the Korean Spotlight version did on a first glance not lead to worse results when comparing with the English version of Spotlight. As previously mentioned, we expect POS tags to be of higher importance to processing speed.

The quality of the input knowledge base (DBpedia, Wikipedia) has an unexpected result on the system performance. Though the English DBpedia contains a much higher number of resources and triples, the Korean DBpedia Spotlight approach outperformed the English DBpedia Spotlight. The performance differences from English to Korean Spotlight approaches can only be explained with differences in the knowledge source. It becomes obvious that a large set of candidate entities may lead to a worse performance if measured by precision, recall and f-measure. We believe this is caused by the fact that resources in the English DBpedia not existent in the Korean DBpedia mostly describe entities that are rarely occurring in input text. Thus, they should be treated with more attention during disambiguation. The influence of the size and quality on a system's performance is thus tied to the used algorithms for semantic annotation. The Korean Spotlight only has access to more frequent entities, simplifying the disambiguation process and reducing the risk of

Table 5 Fraction of retrievable entities with M&W's tokenization approach

	English	Korean
Fraction of retrievable entities	1.000	0.193

wrong disambiguation decisions. If the goal of an annotation system is to achieve a maximum in precision, depending on the system's annotation logic it is likely that an optimal size of an input knowledge source exists, which is somewhere between 0 and the size of the English DBpedia / English Wikipedia. It might however be argued that retrieving and correctly linking specific entities with few occurrences adds more value to semantic annotation than the annotation of frequently occurring general entities. In that case systems would need to be evaluated in a context where manually annotated entities are assigned weights and the retrieval of less general entities results in higher scores. In the current evaluation approaches, each entity annotation in the gold standard is treated equal.

7 Summary and conclusion

This work has identified common challenges existing to the internationalization of semantic annotation systems in the LOD context, adapted two systems for Korean language use, and evaluated system performance on a manually annotated bilingual corpus. The language support of existing systems was found to be still limited with no existing system supporting the Korean language.

Difficulties in the internationalization and localization of semantic annotation systems were found to be caused by applied language processing technology either not being available for a set of new input languages or only partly applicable, or caused by the quality of knowledge bases related to its size, structure of links between entities, semantic richness of information attached to entities, and the availability of lexicalization data.

Candidate systems for adaption to the Korean language were identified to be the DBpedia Spotlight approach [15] and Milne and Witten's approach [17], each of which provides open access to source code. The two previously mentioned approaches were adapted to Korean language use without major changes in the system logic and evaluated on a bilingual sense-tagged corpus resource that we created within this research.

Following this research, further work will be directed towards identifying the root causes of language-specific weaknesses shown in our evaluation. The work we have completed so far intends to be a first step in the direction of developing a fully functional solution for semantic annotation of Korean language text with LOD resources and a general guideline for the internationalization of semantic annotation approaches.

Acknowledgements This research was conducted by the International Collaborative Research and Development Program (Creating Knowledge out of Interlinked Data) and funded by the Korean Ministry of Knowledge Economy.

References

1. Auer S, Bizer C, Kobilarov G, Lehmann J (2007) DBpedia: a nucleus for a web of open data. In: 6th international semantic web conference (ISWC07)
2. Auer S, Weidl M, Lehmann J, Zaveri AJ, Choi KS (2010) I18n of semantic web applications. In: 9th international semantic web conference (ISWC10)

3. Benjamins V, Contreras J, Corcho O (2002) Six challenges for the semantic web. In: KR2002 workshop on formal ontology, knowledge representation and intelligent systems for the web
4. Chai H (2007) Automatic annotation for korean - approach based on the contextual exploration method. In: Database and expert systems applications (DEXA07)
5. Chai H, Djioua B, Le Priol F (2010) Korean semantic annotation on the EXCOM platform. In: Proceedings of the 21st Pacific Asia conference on language, information and computation
6. Chung T, Post M (2010) Factors affecting the accuracy of korean parsing. In: NAACL HLT 2010 first workshop on statistical parsing of morphologically-rich languages (SPMRL10)
7. Djioua B, Flores J, Blais A, Desclés J (2006) EXCOM: an automatic annotation engine for semantic information. In: Proceedings of the FLAIRS conference 2006
8. Ferragina P (2010) TAGME: on-the-fly annotation of short text fragments (by Wikipedia Entities). In: 19th ACM conference on information and knowledge management (CIKM10)
9. Gerber A, Gao L (2011) A scoping study of (who, what, when, where) semantic tagging services. University of Queensland, Australia
10. Halpern J (2006) The contribution of lexical resources to natural language processing of CJK languages. In: 5th international conference on chinese spoken language processing (ISCSLP06)
11. Heath T, Bizer C (2011) Linked data: evolving the web into a global data space. In: Synthesis lectures on the semantic web
12. Kim E, Weidl M, Choi K (2010) Towards a Korean DBpedia and an approach for complementing the Korean Wikipedia based on DBpedia. In: Proceedings of the 5th open knowledge conference
13. Medelyan O, Witten I (2008) Topic indexing with Wikipedia In: Proceedings of the AAAI WikiAI workshop
14. Meij E, Weerkamp W (2012) Adding semantics to microblog posts. In: 5th ACM international conference on web search and data mining (WSDM12)
15. Mendes P, Jakob M, Garcia-Silva A (2011) DBpedia spotlight: shedding light on the web of documents. In: 7th international conference on semantic systems (I-Semantics)
16. Mihalcea R (2007) Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on information and knowledge management (CIKM2007)
17. Milne D (2008) Learning to link with Wikipedia. In: 17th ACM conference on information and knowledge management (CIKM08)
18. Milne D (2009) An open-source toolkit for mining Wikipedia. In: Proceedings of New Zealand computer science research
19. Ratinov L, Roth D, Downey D (2011) Local and global algorithms for disambiguation to Wikipedia. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies (HLT2011)
20. Rizzo G (2011) NERD: evaluating named entity recognition tools in the web of data. In: 10th international semantic web conference (ISWC2011)
21. Zheng H, Kang B, Koo S, Choi H (2006) A semantic annotation tool to extract instances from korean web documents. In: 1st semantic authoring and annotation workshop of 5th international semantic web conference (ISWC2006)

David Müller is a recent graduate of the Department of Business Engineering, Karlsruhe Institute of Technology (KIT), Germany. He spent seven months in 2011/2012 at the Department of Knowledge Service Engineering at KAIST, Korea, combining his passion for semantic technologies and Korean language in his research on semantic annotation of multilingual resources under the supervision of Professor Mun Yong Yi as a part of KAIST's involvement in the Linked Open Data 2 (www.lod2.eu) project. He has further been participating in research in the fields of Semantic Web and Linked Data at the Institute of Applied Informatics and Formal Description Methods (AIFB) at the Karlsruhe Institute of Technology. David Mueller's current research interests include large-scale data processing, distributed machine learning, knowledge mining and cross-language entity linking.

Mun Yong Yi is an Associate Professor in the Department of Knowledge Service Engineering and the director of Knowledge Systems Lab at KAIST. He is participating in the Linked Open Data (http://lod2.edu) project, in which several European research institutions and KAIST collaborate on the development of tools for the Semantic Web. He earned his Ph.D. in Information Systems from University of Maryland, College Park. Before joining KAIST in 2009, he taught at University of South Carolina as an Assistant Professor (1998–2004) and (tenured) Associate Professor (2005–2009). His current research interests include semantic information retrieval, recommender systems, knowledge structure engineering, computer skill acquisition, and technology adoption. His work has been published in a number of journals including *Information Systems Research*, *Decision Sciences*, *Information & Management*, *International Journal of Human-Computer Studies*, and *Journal of Applied Psychology*. He is a former editorial member of *MIS Quarterly* and a current Associate Editor for *International Journal of Human-Computer Studies* and a Senior Editor for *AIS Transactions on Human-Computer Interaction*.