Cross-lingual Alignment and Completion of Wikipedia Templates

Gosse Bouma
Information Science
University of Groningen
g.bouma@rug.nl

Sergio Duarte
Information Science
University of Groningen
sergio.duarte@gmail.com

Zahurul Islam
Information Science
University of Groningen
zaisdb@gmail.com

Abstract

For many languages, the size of Wikipedia is an order of magnitude smaller than the English Wikipedia. We present a method for cross-lingual alignment of template and infobox attributes in Wikipedia. The alignment is used to add and complete templates and infoboxes in one language with information derived from Wikipedia in another language. We show that alignment between English and Dutch Wikipedia is accurate and that the result can be used to expand the number of template attribute-value pairs in Dutch Wikipedia by 50%. Furthermore, the alignment provides valuable information for normalization of template and attribute names and can be used to detect potential inconsistencies.
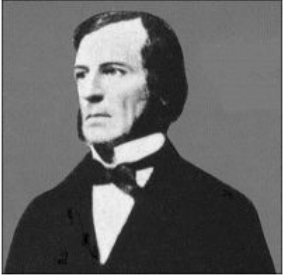
1 Introduction

One of the more interesting aspects of Wikipedia is that it has grown into a multilingual resource, with Wikipedia's for many languages, and systematic (cross-language) links between the information in different language versions. Eventhough English has the largest Wikipedia for any given language, the amount of information present in Wikipedia exceeds that of any single Wikipedia. One of the reasons for this is that each language version of Wikipedia has its own cultural and regional bias. It is likely, for instance, that information about the Netherlands is better represented in Dutch Wikipedia than in other Wikipedia's. Some indication that this is indeed the case comes from the fact a Google search for *'Pim Fortuyn'* in the Dutch Wikipedia gives 498 hits, whereas the English Wikipedia gives only 292 hits. Also, 21,697 pages in Dutch Wikipedia fall in a category matching *'Nederlands(e)'*, whereas only 9,494 pages in English Wikipedia fall in a category matching *'Dutch'*. This indicates that, apart from the obvious fact that smaller Wikipedia's can be expanded with information found in the larger Wikipedia's, it is also true that even the larger Wikipedia's can be supplemented with information harvested from smaller Wikipedia's.

Wikipedia infoboxes are tabular summaries of the most relevant facts contained in an article. They represent an important source of information for general users of the encyclopedia. Infoboxes (see figure 1) encode facts using attributes and values, and therefore are easy to collect and process automatically. For this reason, they are extremely valuable for systems that harvest information from Wikipedia automatically, such as DbPedia (Auer et al., 2008). However, as Wu and Weld (2007) note, infoboxes are missing for many pages, and not all infoboxes are complete. This is particularly true for Wikipedia's in languages other than English.

Infoboxes are a subclass of Wikipedia templates, which are used by authors of Wikipedia pages to express information in a systematic way, and to ensure that formatting of this information is consistent across Wikipedia. Templates exist for referring to multimedia content, external websites, news stories, scientific sources, other on-line repositories (such as the Internet Movie Database (IMDB), medical classification systems (ICD9 and ICD10), coordinates on Google Maps, etc. Although we are primarily interested in infoboxes, in the experiments below we take

Western Philosophy
19th-century philosophy

George Boole

Full name	George Lawlor Boole
Birth	November 2, 1815 (Lincoln Lincolnshire , England)
Death	December 8, 1864 (aged 49) (Ballintemple, County Cork, Ireland),(Drug Overdose)
School/tradition	Mathematical foundations of computer science
Main interests	Mathematics, Logic, Philosophy of mathematics
Notable ideas	Boolean algebra

```
{{Infobox Philosopher |
 region = Western Philosophy |
 era = [[19th-century philosophy]] |
 image_name = George Boole.jpg|
 image_caption = George Boole |
 name = George Lawlor Boole |
 birth = November 2, 1815 |
 death = December 8, 1864 |
 school = Mathematical foundations
         of [[computer science]] |
 main_interests = [[Mathematics]], [[Logic]] |
 ideas = [[Boolean algebra]]
}}
```

Figure 1: Infobox and (simplified) Wikimedia source

all templates into account.

We plan to use information mined from Wikipedia for Question Answering and related tasks. In 2007 and 2008, the CLEF question answering track[1] used Wikipedia as text collection. While the usual approach to open domain question answering relies on information retrieval for selecting relevant text snippets, and natural language processing techniques for answer extraction, an alternative stream of research has focussed on the potential of on-line data-sets for question answering (Lita et al., 2004; Katz et al., 2005). In Bouma et al. (2008) it is suggested that information harvested from infoboxes can be used for question answering in CLEF. For instance, the answer to questions such as *How high is the Matterhorn?, Where was Piet Mondriaan born?*, and *What is the area of the country Suriname?* can in principle be found in infoboxes. However, in practice the number of questions that is answered by their Dutch QA-system by means information from infoboxes is small. One reason for this is the lack of coverage of infoboxes in Dutch Wikipedia.

In the recent GIKICLEF task[2] systems have to find Wikipedia pages in a number of languages which match descriptions such as *Which Australian mountains are higher than 2000 m?, French bridges which were in construction between 1980 and 1990,* and *African capitals with a population of two million inhabitants or more.* The emphasis in this task is less on answer extraction from text (as in QA) and more on accurate interpretation of (geographical) facts known about an entity. GIKICLEF is closely related to the *entity ranking* task for Wikipedia, as organized by INEX.[3] We believe systems participating in tasks like this could profit from large collections of ⟨*entity,attribute,value*⟩ triples harvested from Wikipedia templates.

In this paper, we propose a method for automatically expanding the amount of information present in the form of templates. In our experiments, we used English and Dutch Wikipedia as sources. Given a page in English, and a matching page in Dutch, we first find all English-Dutch attribute-value

---

[1] http://clef-qa.itc.it

[2] http://www.linguateca.pt/GikiCLEF
[3] http://inex.is.informatik.uni-duisburg. de

tuples which have a matching value. Based on the frequency with which attributes match, we create a bidirectional, intersective, alignment of English-Dutch attribute pairs. Finally, we use the set of aligned attributes to expand the number of attribute-value pairs in Dutch Wikipedia with information obtained from matching English pages. We also show that aligned attributes can be used to normalize attribute names and to detect formatting issues and potential inconsistencies in attribute values.

2 Previous Work

DbPedia (Auer et al., 2008) is a large, on-going, project which concentrates on harvesting information from Wikipedia automatically, on normalization of the extracted information, on linking the information with other on-line data repositories, and on interactive access. It contains 274M facts about 2.6M entities (November, 2008). An important component of DbPedia is harvesting of the information present in infoboxes. However, as Wu and Weld (2007) note, not all relevant pages have (complete) infoboxes. The information present in infoboxes is typically also present in the running text of a page. One line of research has concentrated on using the information obtained from infoboxes as seeds for systems that learn relation extraction patterns (Nguyen et al., 2007). Wu and Weld (2007) go one step further, and concentrate on learning to complete the infoboxes themselves. They present a system which first learns to predict the appropriate infobox for a page (using text classification). Next, they learn relation extraction patterns using the information obtained from existing infoboxes as seeds. Finally, the learned patterns are applied to text of pages for which a new infobox has been predicted, to assign values to infobox attributes. A recent paper by Adar et al. (2009) (which only came to our attention at the time of writing) starts from the same observation as we do. It presents a system for completing infoboxes for English, German, French, and Spanish Wikipedia, which is based on learning a mapping between infoboxes and attributes in multiple languages. A more detailed comparison between their approach and ours is given in section 6

The potential of the multilingual nature of Wikipedia has been explored previously by several researchers. Adafre and de Rijke (2006) explore machine translation and (cross-lingual) link structure to find sentences in English and Dutch Wikipedia which express the same content. Bouma et al. (2006) discuss a system for the English-Dutch QA task of CLEF. They basically use a Dutch QA-system, which takes questions automatically translated from English (by the on-line Babelfish translation service). To improve the quality of the translation of named entities, they use, among others, cross-language links obtained from Wikipedia. Erdmann et al. (2008) explore the potential of Wikipedia for the extraction of bilingual terminology. They note that apart from the cross-language links, page redirects and anchor texts (i.e. the text that is used to label a hypertext reference to another (wikipedia) page) can be used to obtain large and accurate bilingual term lists.

3 Data collection and Preparation

We used a dump of Dutch Wikipedia (June 2008) and English Wikipedia (August 2007) made available by the University of Amsterdam[4] and converted to an XML-format particularly suitable for information extraction tasks.

From these two collections, for each page, we extracted all attribute-value pairs found in all templates. Results were stored as quadruples of the form ⟨*Page, TemplateName, Attribute, Value*⟩. Each *TemplateName~Attribute* pair expresses a specific semantic relation between the entity or concept described by the *Page* and a *Value*. Values can be anything, but often refer to another Wikipedia page (i.e. ⟨*George Boole, Philosopher, notable_ideas, Boolean algebra*⟩, where *Boolean algebra* is a link to another page) or to numeric values, amounts, and dates. Note that attributes by themselves tend to be highly ambiguous, and often can only be interpreted in the context of a given template. The attribute *period*, for instance, is used to describe chemical elements, royal dynasties, countries, and (historical) means of transportation. Another source of attribute ambiguity is the fact that many templates simply number their attributes. As we are interested in finding an alignment between semantically meaningful relations in the two collections, we will therefore

---

[4] `http://ilps.science.uva.nl/WikiXML`

	Dutch	English
date	June 2008	August 2007
pages	715,992	3,840,950
pages with template	290,964	757,379
cross-language links		126,555
templates	550,548	1,074,935
tuples	4,357,653	5,436,033
template names	2,350	7,783
attribute names	7,510	19,378
templ∼attr pairs	23,399	81,671

Table 1: Statistics for the version of Dutch and English Wikipedia used in the experiment.

concentrate on the problem of finding an alignment between *TemplateName∼Attribute* pairs in English and Dutch Wikipedia.

Some statistics for the two collections are given in table 1. The number of pages is the count for all pages in the collection. It should be noted that these contain a fair number of administrative pages, pages for multimedia content, redirect pages, page stubs, etc. The number of pages which contains content that is useful for our purposes is therefore probably a good deal lower, and is maybe closer to the number of pages containing at least one template. Cross-language links (i.e. links from an English page to the corresponding Dutch page) where extracted from English Wikipedia. The fact that 0.5M templates in Dutch give rise to 4.3M tuples, whereas 1.0M templates in English give rise to only 5.4M tuples is perhaps a consequence of the fact that the two collections are not from the same date, and thus may reflect different stages of the development of the template system.

We did spend some time on normalization of the values found in extracted tuples. Our alignment method relies on the fact that for a sufficient number of matching pages, tuples can be found with matching values. Apart from identity and Wikipedia cross-language links, we rely on the fact that dates, amounts, and numerical values can often be recognized and normalized easily, thus increasing the number of tuples which can be used for alignment. Normalization addresses the fact that the use of comma's and periods (and spaces) in numbers is different in English and Dutch Wikipedia, and that

dates need to be converted to a standard. English Wikipedia expresses distances and heights in miles and feet, weights in pounds, etc., whereas Dutch Wikipedia uses kilometres, metres, and kilograms. Where English Wikipedia mentions both miles and kilometres we preserve only the kilometres. In other situations we convert miles to kilometres. In spite of this effort, we noted that there are still quite a few situations which are not covered by our normalization patterns. Sometimes numbers are followed or preceded by additional text (*approx.14.5 MB, 44 minutes per episode*), sometimes there is irregular formatting (*October 101988*), and some units simply are not handled by our normalization yet (i.e. converting square miles to square kilometres). We come back to this issue in section 6.

Krötzsch et al. (2007) have aslo boserved that there is little structure in the way numeric values, units, dates, etc. are represented in Wikipedia. They suggest a tagging system similar to the way links to other Wikipedia pages are annotated, but now with the aim of representing numeric and temporal values systematically. If such a system was to be adopted by the Wikipedia community, it would greatly facilitate the processing of such values found in infoboxes.

4 Alignment

In this section we present our method for aligning English Template∼Attribute pairs with corresponding pairs in Dutch.

The first step is creating a list of matching tuples.

> **Step 1.** Extract all matching template tuples.
> An English $\langle Page_e,\ Templ_e{\sim}Attr_e,\ Val_e \rangle$ tuple matches a Dutch $\langle Page_d,\ Templ_d{\sim}Attr_d,\ Val_d \rangle$ tuple if $Page_e$ matches $Page_d$ and Val_e matches Val_d and there is no other tuple for either $Page_e$ or $Page_d$ with value Val_e or Val_d.
> Two pages or values E and D match if there exists a cross-language link which links E and D, or if $E{=}D$.

We only take into account tuples for which there is a unique (non-ambiguous) match between English and Dutch. Many infoboxes contain attributes which

often take the same value (i.e. *title* and *imdb_title* for movies). Other cases of ambiguity are caused by numerical values which incidentally may take on identical values. Such ambiguous cases are ignored. Step 1 gives rise to 149,825 matching tuples.[5] It might seem that we find matching tuples for only about 3-4% of the tuples present in Dutch Wikipedia. Note, however, that while there are 290K pages with a template in Dutch Wikipedia, there are only 126K cross-language links. The total numer of tuples on Dutch pages for which a cross-language link to English exists is 837K. If all these tuples have a counterpart in English Wikipedia (which is highly unlikely), our method finds a match for 18% of the relevant template tuples.[6]

The second step consists of extracting matching English-Dutch Template∼Attribute pairs from the list of matching tuples constructed in step 1.

> **Step 2.** For each matching pair of tuples $\langle Page_e, Templ_e{\sim}Attr_e, Val_e\rangle$ and $\langle Page_d, Templ_d{\sim}Attr_d, Val_d\rangle$, extract the English-Dutch pair of Template∼Attributes $\langle Templ_e{\sim}Attr_e, Templ_d{\sim}Attr_d\rangle$.

In total, we extracted 7,772 different English-Dutch $\langle Templ_e{\sim}Attr_e, Templ_d{\sim}Attr_d\rangle$ tuples. In 547 cases $Templ_e{\sim}Attr_e = Templ_d{\sim}Attr_d$. In 915 cases, $Attr_e = Attr_d$. In the remaining 6,310 cases, $Attr_e \neq Attr_d$. The matches are mostly accurate. We evaluated 5% of the matching template∼attribute pairs, that had been found at least 2 times. For 27% of these (55 out 205), it was not immediately clear whether the match was correct, because one of the attributes was a number. Among the remaining 150 cases, the only clear error seemed to be a match between the attributes *trainer* and *manager* (for soccer club templates). Other cases which are perhaps not always correct were mappings between *successor, successor1, successor2* on the one hand and *after/next* on the other hand. The attributes with a

---

[5]It is interesting to note that 51K of these matching tuples are for pages that have an identical name in English and Dutch, but were absent in the table of cross-language links. As a result, we find 32K pages with an identical name in English and Dutch, and at least one pair of matching tuples. We suspect that these newly discovered cross-language links are highly accurate.

[6]If we also include English pages with a name identical to a Dutch page, the maximum number of matching tuples is 1.1M, and we find a match for 14% of the data

101	cite_web	title
27	voetnoot_web	titel
12	film	titel
10	commons	1
7	acteur	naam
6	game	naam
5	ster	naam
4	taxobox_zoogdier	w-naam
4	plaats	naam
4	band	band_naam
3	taxobox	w-naam

Table 2: Dutch template∼attribute pairs matching English *cite_web∼title*. Counts refer to the number of pages with a matching value.

number suffix probably refer to the *nth* successor, whereas the attributes without suffix probably refer to the immediate successor.

On the other hand, for some frequent template∼attribute pairs, ambiguity is clearly an issue. For the English pair *cite_web∼title* for instance, 51 different mappings are found. The most frequent cases are shown in table 2. Note that it would be incorrect to conclude from this that, for every English page which contains a *cite_web∼title* pair, the corresponding Dutch page should include, for instance, a *taxobox∼w-naam* tuple.

In the third and final step, the actual alignment between English-Dutch template∼attribute pairs is established, and ambiguity is eliminated.

> **Step 3.** Given the list of matching template∼attribute pairs computed in step 2 with a frequency ≥ 5, find for each English $Templ_e{\sim}Attr_e$ pair the most frequent matching Dutch pair $Templ_d{\sim}Attr_d$. Similarly, for each Dutch pair $Templ_d{\sim}Attr_d$, find the most frequent English pair $Templ_e{\sim}Attr_e$. Return the intersection of both lists.

2,070 matching template∼attribute tuples are seen at least 5 times. Preference for the most frequent bidirectional match leaves 1,305 template∼attribute tuples. Examples of aligned tuples are given in table 3. We evaluated 10% of the tuples containing meaningful attributes (i.e. not

	English		Dutch	
	Template	Attribute	Template	Attribute
	actor	spouse	acteur	partner
	book	series	boek	reeks
	casino	owner	casino	eigenaar
	csi_character	portrayed	csi_personage	acteur
	dogbreed	country	hond	land
	football_club	ground	voetbal_club	stadion
	film	writer	film	schrijver
	mountain	range	berg	gebergte
	radio_station	airdate	radiozender	lancering

Table 3: Aligned template~attribute pairs

	pages	triples
existing pages	50,099	253,829
new cross-links	11,526	43,449
new dutch pages	321,069	1,931,277
total	382,694	2,228,555

Table 4: Newly inferred template tuples

numbers or single letters). In 117 tuples, we discovered two errors: ⟨*aircraft_specification~number of props, gevechtsvliegtuig~bemanning*⟩ aligns the number of engines with the number of crew members (based on 10 matching tuples), and ⟨*book~country, film~land*⟩ involves a mismatch of templates as it links the country attribute for a book to the country attribute for a movie.

Note that step 3 is similar to bidirectional intersective word alignment as used in statistical machine translation (see Ma et al. (2008), for instance). This method is known for giving highly precise results.

5 Expansion

We can use the output of step 3 of the alignment method to check for each English tuple whether a corresponding Dutch tuple can be predicted. If the tuple does not exist yet, we add it. In total, this gives rise to 2.2M new tuples for 382K pages for Dutch Wikipedia (see table 4). We generate almost 300K new tuples for existing Dutch pages (250K for pages for which a cross-language link already existed). This means we exand the total number of tuples for existing pages by 27%. Most tuples, however, are generated for pages which do not yet exist in Dutch Wikipedia. These are perhaps less useful, although one could use the results as knowledge for a QA-system, or to generate stubs for new Wikipedia pages which already contain an infobox and other relevant templates.

The 100 most frequenty added template~attribute pairs (ranging from *music album~genre* (added 31,392 times) to *single~producer* (added 5605

times)) are dominated by templates for music albums, geographical places, actors, movies, and taxonomy infoboxes.

We evaluated the accuracy of the newly generated tuples for 100 random existing Dutch wikipedia pages, to which at least one new tuple was added. The pages contained 802 existing tuples. 876 tuples were added by our automatic expansion method. Of these newly added tuples, 62 contained a value which was identical to the value of an already existing tuple (i.e. we add the tuple ⟨*Reuzenhaai, taxobox~naam, Reuzenhaai*⟩ where there was already an existing tuple ⟨*Reuzenhaai, taxobox_begin~name, Reuzenhaai*⟩ tuple – note that we add a properly translated attribute name, where the original tuple contains a name copied from English!). The newly added tuples contained 60 tuples of the form ⟨*Aegna, plaats~lat_dir, N (letter)*⟩, where the value should have been N (the symbol for latitude on the Northern hemisphere in geographical coordinates), and not the letter N. One page (*Akira*) was expanded with an incoherent set of tuples, based on tuples for the manga, anime, and music producer with the same name. Apart from this failure, there were only 5 other clearly incorrect tuples (adding o.a. *place~name* to *Albinism*, adding *name in Dutch* with an English value to *Macedonian*, and adding *community~name* to *Christopher Columbus*). In many cases, added tuples are based on a different template for the same entity, often leading to almost identical values (i.e. adding geographical coordinates using slightly different notation). In one case, *Battle of Dogger Bank (1915)*, the system added new tuples based on a template that was already in use for the Dutch page as well, thus automatically updating and expanding an existing template.

	geboren		population
23	birth_date	49	inwoners
16	date of birth	9	population
8	date_of_birth	5	bevolking
8	dateofbirth	4	inwonersaantal
2	born	3	inwoneraantal
2	birth	2	town pop
1	date_birth	2	population_total
1	birthdate	1	townpop
		1	inw.
		1	einwohner

Table 6: One-to-many aligned attribute names. Counts are for the number of (aligned) infoboxes that contain the attribute.

6 Discussion

6.1 Detecting Irregularities

Instead of adding new information, one may also search for attribute-value pairs in two Wikipedia's that are expected to have the same value, but do not. Given an English page with attribute-value pair $\langle Attr_e, Val_e \rangle$, and a matching Dutch page with $\langle Attr_d, Val_d \rangle$, where $Attr_e$ and $Attr_d$ have been aligned, one expects Val_e and Val_d to match as well. If this is not the case, something irregular is observed. We have applied the above rule to our dataset, and detected 79K irregularities. An overview of the various types of irregularities is given in table 5. Most of the non-matching values are the result of formatting issues, lack of translations, one value being more specific than the other, and finally, inconsistencies. Note that inconsistencies may also mean that one value is more recent than the other (population, (stadium) capacity, latest release data, spouse, etc.). A number of formatting issues (of numbers, dates, periods, amounts, etc.) can be fixed easily, using the current list of irregularities as starting point.

6.2 Normalizing Templates

It is interesting to note that alignment can also be used to normalize template attribute names. Table 6 illustrates this for the Dutch attribute *geboren* and the English attribute *population*. Both are aligned with a range of attribute names in the other language.

Such information is extremely valuable for applications that attempt to harvest knowledge from Wikipedia, and merge the result in an ontology, or attempt to use the harvested information in an application. For instance, a QA-system that has to answer questions about birth dates or populations, has to know which attributes are used to express this information. Alternatively, one can also use this information to normalize attribute-names. In that case, all attributes which express the birth date property could be replaced by *birth_date* (the most frequent attribute currently in use for this relation).

This type of normalization can greatly reduce the *noisy* character of the current infoboxes. For instance, there are many infoboxes in use for geographic locations, people, works of art, etc. These infoboxes often contain information about the same properties, but, as illustrated above, there is no guarantee that these are always expressed by the same attribute.

6.3 Alignment by means of translation

Template and attribute names in Dutch often are straightforward translations of the English name, e.g. *luchtvaartmaatschappij/airline, voetbalclub/football club, hoofdstad/capital, naam/name, postcode/postalcode, netnummer/area_code* and *opgericht/founded*. One might use this information as an alternative for determining whether two template∼attribute pairs express the same relation.

We performed a small experiment on infoboxes expressing geographical information, using Wikipedia cross-language links and an on-line dictionary as multilingual dictionaries. We found that 10 to 15% of the attribute names (depending on the exact subset of infoboxes taken into consideration) could be connected using dictionaries. When combined with the attributes found by means of alignment, coverage went up to maximally 38%.

6.4 Comparison

It is hard to compare our results with those of Adar et al. (2009). Their method uses a Boolean classifier which is trained using a range of features to determine whether two values are likely to be equivalent (including identity, string overlap, link relations, translation features, and correlation of numeric values). Training data is collected automatically by

Attributes		Values		Type
English	Dutch	English	Dutch	
capacity	capaciteit	23,400	23 400	formatting
nm	lat_min	04	4	formatting
date	date	1775-1783	1775–1783	formatting
name	naam	African baobab	Afrikaanse baobab	translation
artist	artiest	Various Artists	Verschillende artiesten	translation
regnum	rijk	Plantae	Plantae (Planten)	specificity
city	naam,	Comune di Adrara San Martino	Adrara San Martino	specificity
birth_date	geboren	1934	1934-8-25	specificity
imagepath_coa	wapenafbeelding	coa_missing.jpg	Alvaneu wappen.svg	specificity
population	inwonersaantal	5345	5369	inconsistent
capacity	capaciteit	13,152	14 400	inconsistent
dateofbirth	geboortedatum	2 February 1978	1978-1-2	inconsistent
elevation	hoogte	300	228	inconsistent

Table 5: Irregular values in aligned attributes on matching pages

selecting highly similar tuples (i.e. with identical template and attribute names) as positive data, and a random tuple from the same page as negative data. The accuracy of the classifier is 90.7%. Next, for each potential pairing of template~attribute pairs from two languages, random tuples are presented to the classifier. If the ratio of positively classified tuples exceeds a certain threshold, the two template~attribute pairs are assumed to express the same relation. The accuracy of result varies, with matchings of template~attribute pairs that are based on the most frequent tuple matches having an accuracy score of 60%. They also evaluate their system by determining how well the system is able to predict known tuples. Here, recall is 40% and precision is 54%. The recall figure could be compared to the 18% tuples (for pages related by means of a cross-language link) for which we find a match. If we use only properly aligned template~attribute pairs, however, coverage will certainly go down somewhat. Precision could be compared to our observation that we find 149K matching tuples, and, after alignment, predict an equivalence for 79K tuples which in the data collecting do not have a matching value. Thus, for 228K tuples we predict an equivalent values, whereas this is only the case for 149K tuples. We would not like to conclude from this, however, that the precision of our method is 65%, as we observed in section 5 that most of the conflicting values are not inconsistencies, but more often the consequence of formatting irregularities, transla-

tions, variation in specificity, etc. It is clear that the system of Adar et al. (2009) has a higher recall than ours. This appears to be mainly due to the fact that their feature based approach to determining matching values considers much more data to be equivalent than our approach which normalizes values and then requires identity or a matching cross-language link. In future work, we would like to explore more rigorous normalization (taking the data discussed in section 5 as starting point) and inclusion of features to determine approximate matching to increase recall.

7 Conclusions

We have presented a method for automatically completing Wikipedia templates which relies on the multilingual nature of Wikipedia and on the fact that systematic links exist between pages in various languages. We have shown that matching template tuples can be found automatically, and that an accurate set of matching template~attribute pairs can be derived from this by using intersective bidirectional alignment. The method extends the number of tuples by 51% (27% for existing Dutch pages).

In future work, we hope to include more languages, investigate the value of (automatic) translation for template and attribute alignment, investigate alternative alignment methods (using more features and other weighting scheme's), and incorporate the expanded data set in our QA-system for Dutch.

References

S.F. Adafre and M. de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

E. Adar, M. Skinner, and D.S. Weld. 2009. Information arbitrage across multi-lingual Wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 94–103. ACM New York, NY, USA.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2008. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735.

Gosse Bouma, Ismail Fahmi, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jörg Tiedemann. 2006. The University of Groningen at QA@CLEF 2006: Using syntactic knowledge for QA. In *Working Notes for the CLEF 2006 Workshop*, Alicante.

Gosse Bouma, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jörg Tiedemann. 2008. Question answering with Joost at QA@CLEF 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus.

M. Erdmann, K. Nakayama, T. Hara, and S. Nishio. 2008. An approach for extracting bilingual terminology from wikipedia. *Lecture Notes in Computer Science*, 4947:380.

B. Katz, G. Marton, G. Borchardt, A. Brownell, S. Felshin, D. Loreto, J. Louis-Rosenberg, B. Lu, F. Mora, S. Stiller, et al. 2005. External knowledge sources for question answering. In *Proceedings of the 14th Annual Text REtrieval Conference (TREC'2005), November*.

M. Krötzsch, D. Vrandečić, M. Völkel, H. Haller, and R. Studer. 2007. Semantic wikipedia. *Web Semantics: Science, Services and Agents on the World Wide Web*.

L.V. Lita, W.A. Hunt, and E. Nyberg. 2004. Resource analysis for question answering. In *Association for Computational Linguistics Conference (ACL)*.

Y. Ma, S. Ozdowska, Y. Sun, and A. Way. 2008. Improving word alignment using syntactic dependencies. In *Proceedings of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, pages 69–77.

D.P.T. Nguyen, Y. Matsuo, and M. Ishizuka. 2007. Relation extraction from wikipedia using subtree mining. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, page 1414. AAAI Press.

Fei Wu and Daniel S. Weld. 2007. Autonomously semantifying wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50, New York, NY, USA. ACM.