Extracting and Mining Cross-Lingual Wikipedia

Michael Skinner & Eytan Adar

CSE 574 Project Proposal

Though the English subdomain of Wikipedia is ranked #1 in page counts with 2.1 million articles, this represents only 22% of Wikipedia. The remaining 78% of effort is distributed among over 250 languges (though principally in the top 50 or so). As Wikipedians rush to create new content and translate and fill in the gaps in various languages there is a tremendous opportunity to apply automated extraction techniques in the space. In particular, many topics are maintained more effectively by editors in one language over all others. This may be because a particularly motivated editor only writes in one language and not another. It may also be do due to distribution and availability of certain expertise and information. For example, a French rock-star may have more loyal fans updating the star's page in French giving it a more up-to-date and a more complete article than the page in another language. The differences in the amount, quality, and recency of information in Wikipedia depending on language represents a particular problem for its users, but is also an interesting opportunity and challenge for us.

For our 574 project, we propose to create a system that extracts information in one language for filling in gaps or update facts in another. In particular, we hope to concentrate on infoboxes, tables, and lists as they represent a more manageable first step that does not require deep linguistic understanding or translation. We propose the following parts to the system:

- **Cross-lingual Alignment**: Though many Wiki pages have (human-generated) links to matching articles in other language domains, this is not a consistent or required part of any article. Further, because articles can be split into sub-articles, frequently sub-articles will not contain links. The first part of the problem then is to perform an alignment between the Wikipedia graph in one language to that of the other. We propose concentrating on 3-4 large language domains to start (English, and likely some subset of German, French, Spanish, or Portugese). Evaluation of this step can be achieved by ignoring the manually created cross-domain links and attempting to recreate them (the particular amount depends on the number of user-generated links used to for training).

- **Table Mapping (non-Infobox)**

1. **Table Alignment**: Many articles contain lists and tables that hold tabular information which is maintained independently by different editors. Depending on the information that these authors have, different tables may fall out-of sync. The first step in bringing them together is to find tables that in one language that can be mapped to the table in the other. Though this is difficult for tables that are translated, it may be possible to utilize n-gram, sizing, and simple translation strategies (e.g. using the pan-lingual TransGraph for word-by-word translation) to identify related tables. Evaluation will require some manual labeling and comparison, though simulated data sets can be created semi-automatically by translating a table into a different language, randomly numerical data, and re-ordering tables.
2. **Table Synchronization**: Once related tables are aligned, the next step is to align specific entries. This may require sorting in different ways and performing sequence alignment style steps to identify missing and changed information. Once achieved, we

hope to propose reasonable edits for one table based on the other. Again, evaluation can be achieved through a manual process as well as through synthetic, simulated data

- **Information Extraction**

1. **Infobox Alignment and Synchronization:** The same table alignment tasks that worked in non-infobox domain may be used for aligning infoboxes. An interesting side-effect of this alignment is that a glossary can be constructed mapping infobox schemas between languages. Synchronization can be automated or semi-automated, offering a notice to the user that information in the other language is newer (or is available for missing information).
2. **Kylin style extractions (If time allows)**: Because Kylin style extractions are fairly domain independent, it may be interesting to attempt extractions in other languages. The only language dependent feature is the POS tagger, and we may be able to find or train one for at least one other language. The same comparison techniques used on infoboxes and tables for alignment could be used to align Kylin extractions (either with existing infobox entries or other extractions) to improve confidence. Perhaps instead of trying to translate infobox entries between languages, an infobox in one language and a rough translation could be used to create or enhance extraction rules in another.

- **Linked Editing** (**If time allows**): Create an architecture for maintaining links between aligned information. Edits to one can be automatically propagated to the other page, possibly subject to human approval. Even if not part of our project, this is a very promising area for future work.

Each of the components is fairly independent of the others, except for linked editing which requires that there first be some content linked. For the purposes of this project, it is probably best to work on the cross-lingual alignment first, as a high success rate at that would greatly ease the burden associated with finding tables and infoboxes to align. Although baseline functionality is a first goal, work on this component will likely continue throughout the course of the project, as information from the later stages can be used to improve it, either algorithmic improvements to correct flaws that become more apparent in later stages, or something more automatic such as allowing table or infobox match qualities from the later steps to feed back and influence page alignment. After this, techniques for both table and infobox alignment could be developed in parallel. Parallel development helps exploit the similarity of the two tasks, while also aiding in the creation of an infrastructure that is flexible enough to handle both. If time becomes a problem, then one of these two chunks could be dropped, and if the page alignment performs poorly then tables for alignment can be hand picked. If, however, we still have time left over, we could toy with the creation of an architecture that propagates changes between languages based on the linked information discovered in previous stages.

Rough milestones:

- **Feb 13:** Working page alignment -- graph alignment (Michael) based on page similarities (Eytan)
- **Feb 27:** Information alignment -- find similar tables/infoboxes, find mapping between elements, infobox schema mapping, refine page alignment
- **Mar 12:** Improve everything -- refine information alignment and page alignment, explore "bonus" parts of project (Kylin, linked editing), preliminary results and report
- **Mar 21:** Tidying up -- tie up loose ends, finalize results, finish report