MENTA: Inducing Multilingual Taxonomies from Wikipedia

Gerard de Melo
Max Planck Institute for Informatics
Saarbrücken, Germany
demelo@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

In recent years, a number of projects have turned to Wikipedia to establish large-scale taxonomies that describe orders of magnitude more entities than traditional manually built knowledge bases. So far, however, the multilingual nature of Wikipedia has largely been neglected. This paper investigates how entities from all editions of Wikipedia as well as WordNet can be integrated into a single coherent taxonomic class hierarchy. We rely on linking heuristics to discover potential taxonomic relationships, graph partitioning to form consistent equivalence classes of entities, and a Markov chain-based ranking approach to construct the final taxonomy. This results in MENTA (Multilingual Entity Taxonomy), a resource that describes 5.4 million entities and is presumably the largest multilingual lexical knowledge base currently available.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms

1. INTRODUCTION

Motivation. Capturing knowledge in the form of machine-readable semantic knowledge bases has been a long-standing goal in computer science, information science, and knowledge management. Such resources have facilitated tasks like query expansion [20], semantic search [27], faceted search [4], question answering [35], and many more. In the past few years, the open, community-developed encyclopedia Wikipedia has been recognized as a valuable source of such knowledge. Projects like DBpedia [3], YAGO [39], Intelligence-in-Wikipedia [44], and Freebase (*freebase.com*) have exploited the semi-structured nature of Wikipedia to produce valuable repositories of formal knowledge that are orders of magnitude larger than hand-crafted resources like SUMO [29], OpenCyc (*opencyc.org*), or WordNet [17].

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

To date, however, these extraction efforts have largely neglected the significant potential of Wikipedia's multilingual nature. While DBpedia and some other knowledge bases do extract abstracts and other information also from non-English versions, the coverage is still restricted to those entities that have a corresponding article in the English Wikipedia. Certainly, the English Wikipedia is by far the most comprehensive version. Yet, its articles make up only 24% among those of the 50 largest Wikipedias. By aggregating from multiple editions of Wikipedia, we are able to construct MENTA – Multilingual Entity Taxonomy – a large-scale taxonomic knowledge base that covers a significantly greater range of entities than existing knowledge bases. Additionally, MENTA enables tasks like semantic search also in languages other than English, for which existing taxonomies are often very limited or entirely non-existent. Finally, we also hope that MENTA will facilitate decidedly multilingual applications like cross-lingual information retrieval [16, 5], machine translation [24], or learning transliterations [32].

Contribution. The main challenge we tackle is aggregating unreliable taxonomic links between entities from different Wikipedias into a single more reliable and coherent taxonomy. At the heart of our approach lies an algorithm that considers sets of weighted statements linking entities to equivalent entities or parent entities. The input to this algorithm is supplied by a set of heuristic linking functions that connect Wikipedia articles, categories, infoboxes, and WordNet synsets from multiple languages. The algorithm produces aggregated rankings of parents that take into account the dependencies between the linked entities. The output for a specific entity is given by the stationary distribution of a Markov chain, in the spirit of PageRank, but adapted to our specific setting. Overall, this leads to MENTA having three major distinguishing properties.

1. Extended Coverage of Entities: The taxonomy draws on all existing editions of Wikipedia and hence includes large numbers of local places, people, products, etc. that are not covered by the English Wikipedia. For example, the Quechua Wikipedia has an article about the Bolivian salt lake Salar de Coipasa, and the Japanese Wikipedia has an article about Italian Parma ham.

2. Ranked Class Information: Individual entities are linked via `instance` statements to classes (e.g. `City`, `Airline company`, etc.) based on information provided by multiple Wikipedia editions, thus exploiting complementary clues from different languages. Even when an English article provides ample information, it is useful to capture that the Colorado River being a river is more salient than it being a border of Arizona.

3. Coherent Taxonomy: While Wikipedia is an excellent source of semi-structured knowledge about entities, it lacks an ontologically organized taxonomy. The category systems of Wikipedia i) fail to distinguish classes from topic labels (Yellowstone National Park is a natural park but not a 'History of Wyoming', Ulm is a city

but not a 'Swabian League'), ii) tend to lack a clear organization especially at the most abstract level, and iii) differ substantially between different languages. A single, more complete yet coherent ontological class hierarchy is obtained by aggregating information from multiple editions of Wikipedia and WordNet.

The resulting taxonomy in MENTA goes beyond what is currently offered by repositories of semantic knowledge. For instance, DBpedia and YAGO do not have a multilingual upper-level ontology. None of the existing taxonomies have managed to accommodate culture-specific entities from non-English Wikipedia editions. Even for those entities that are covered, the DBpedia Ontology provides class information only for around a third. Likewise, in the field of multilingual taxonomies or hierarchically-organized multilingual lexical knowledge bases, our knowledge base surpasses all existing resources in the number of entities described. MENTA is freely available under an open-source license[1].

Overview. Section 2 lays out how information is extracted from Wikipedia and represented in a form amenable to further processing. Section 3 then introduces the heuristics that are used to interlink entities and provide the input for the taxonomy induction step. Section 4 describes the actual algorithm for producing the unified knowledge base with a single taxonomic class hierarchy. Section 5 evaluates this algorithm and the resulting knowledge base. Section 6 describes related knowledge bases and approaches. Finally, Section 7 provides concluding remarks.

2. KNOWLEDGE EXTRACTION

2.1 Representation Model

As entities, we consider both individual entities as well classes. We regard taxonomies as knowledge bases that describe relationships between entities, including but not limited to ontological relationships that yield a hierarchy of entities. A taxonomy of this form could describe the `Mayflower` as an *instance* of a `Ship`, `Ship` as a *subclass* of `Watercraft`, `Watercraft` as a subclass of `Vehicle`, and so on, up to the taxonomy's universal root node, often called `Entity`. More formally, we rely on the following definitions.

Definition 1. A *statement* is an item from $\mathcal{U} \times \mathcal{R} \times \mathcal{U} \times \mathbb{R}_0^+$, where $\mathcal{U}$ is a universal set of entity identifiers and $\mathcal{R}$ is a set of relations. A statement (x, r, y, w) expresses that two entities x, y stand in relation r to each other with weight w, where a weight of 0 means there is no evidence, and strictly positive values quantify the degree of confidence in the statement being true.

Definition 2. A *knowledge base* K is a tuple $(\mathcal{U}, \mathcal{R}, \mathcal{S})$, where $\mathcal{U}$ is a set of (arbitrary) entity identifiers, $\mathcal{R}$ is a set of relations, and $\mathcal{S}$ is a set of statements that describe relationships between entities.

In our case, $\mathcal{U}$ will contain entity identifiers for Wikipedia pages (including categories and infobox templates), word senses ("synsets") taken from the WordNet database [17], as well as string literals with language designators. The set $\mathcal{R}$ includes:

- `equals`: identity of entities (i.e. two entity identifiers refer to the same entity)
- `subclass`: the relation between a class and a subsuming parent class
- `instance`: the relation between an individual entity and a class it is a member of (its type)
- `means`: the meaning relationship between a language-specific string entity (a word or a name) and another entity

[1] `http://www.mpii.de/yago-naga/menta/`

A statement might express that the `Mayflower` stands in an `instance` relation to the class `Ship` with confidence 1, or that the Czech name 'Curych' stands in a `means` relation to the city of Zürich. Such statements can easily be cast into an RDF [21] form, if reification is used to capture the confidence values.

2.2 Extraction from Wikipedia

Entities. Before aggregating information, we parse the raw XML and wiki-markup-based Wikipedia dumps, extract relevant information, and cast it into our representation model to facilitate further processing. In particular, each article page (including redirect pages), category page, or template page in an edition of Wikipedia is given a preliminary entity identifier. Unfortunately, not all information necessary for parsing the Wikipedia dumps is available from within the dumps alone. We additionally query the web services provided by each server to find out for instance that on the Tagalog Wikipedia, titles starting with "Kategorya:" refer to categories (in addition to the default "Kaurian:" and the English "Category:", which are also accepted). In order to have canonical entity identifiers, such information is normalized.

Statements. Additional information about entities and meta-data about articles that may be of use later on is extracted and stored with appropriate relations, e.g. template invocations, cross-lingual "interwiki" links, and category links. Among other things, we create short descriptions glosses for each article by processing wikitext and HTML mark-up and attempting to identify and possibly truncate the first proper paragraph in an article's wikitext mark-up (ignoring infoboxes, links to disambiguation pages, etc.).

Meanings. Article titles allow us to create `means` statements connecting entities with language-specific strings (labels or names) that refer to them. Some articles use special markup to provide the true capitalization of a title, e.g. 'iPod' instead of 'IPod'. If no markup is provided, we check for the most frequent capitalization variant within the article text.

3. LINKING FUNCTIONS

The first step towards producing a more coherent knowledge base involves exposing connections between different entities. We rely on linking functions to produce statements connecting different entities based on various inputs and heuristics. In particular, Section 3.1 introduces `equals` linking functions that identify identical entities, and Sections 3.2 and 3.3 present linking functions for the `subclass` and `instance` relations. The output of these functions will later serve as input in the actual taxonomy induction step.

Definition 3. Given a relation $r \in \mathcal{R}$, a *linking function* $l_r : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}_0^+$ is a function that yields confidence weight scores $l_r(x, y) \in \mathbb{R}_0^+$ and is used to produce statements $(x, r, y, l_r(x, y))$ for pairs of entities x, y.

Later on,we will explain how information from rather simple, noisy linking functions can be aggregated to provide meaningful results. Hence, the linking functions can make use of heuristics and need not provide perfect results.

3.1 Equality Link Heuristics

We used the following linking functions for evaluating whether `equals` should hold for two entity identifiers x, y.

3.1.1 Cross-Lingual Linking

If x, y are Wikipedia entities connected via cross-lingual interwiki links, e.g. `Zürich` from the English Wikipedia and `Curych` from the Czech one, it returns 1, otherwise 0.

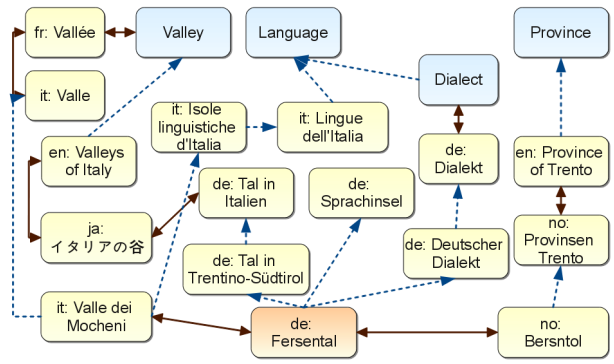

Figure 1: Simplified sample of noisy input from link heuristics

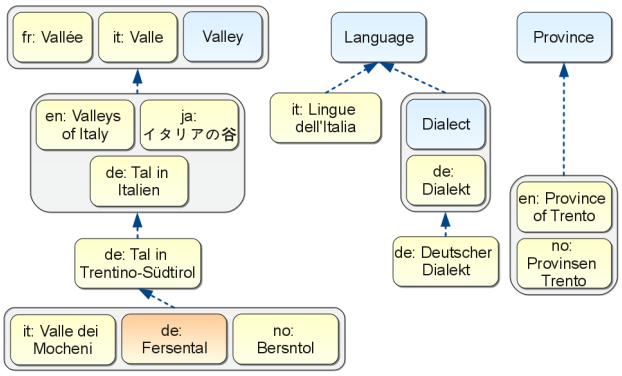

Figure 2: Relevant sample of the desired output

3.1.2 *Category-Article Linking*

This function returns 1 when x, y are a category and an article, respectively, known to be about the same concept, e.g. the category `Abugida writing systems` and the article `Abugida`. This is detected by checking for specific templates on the category page.

3.1.3 *Supervised WordNet Disambiguation*

To see if a Wikipedia article, category, or infobox matches a WordNet synset, we use a linker that assesses their similarity and is trained on a small set of manually labelled training examples to disambiguate possible meanings (see Section 5.2). The linker relies on Ridge Regression [6] to obtain a statistical model from the training examples. As input, it uses three major signals as features.

Term Overlap. The term overlap quantifies the degree of overlap between the respective natural language labels. The set of labels for a Wikipedia entity is given by its title (after capitalization detection) and titles of its redirection articles. A set of labels for a WordNet entity is retrieved from the English, Arabic, Catalan, Estonian, Hebrew, and Spanish wordnets (see *globalwordnet.org*), as well as from MLSN [10]. For a Wikipedia entity x and a WordNet entity y, the term overlap feature is then computed as:

$$\sum_{l_x \in \text{labels}(x)} \max_{l_y \in \text{labels}(y)} w_x(l_x, x) w_y(l_y, y) \text{sim}(l_x, l_y) \qquad (1)$$

Here, $\text{sim}(l_x, l_y)$ is a a simple similarity measure between labels that returns 1 if the languages match and the strings match after lemmatizing and removing additional qualifications in parentheses, and 0 otherwise. For Wikipedia, the additional label weighting w_x generally yields 1, while for WordNet $1/n$ is returned by w_y when n different meanings of l_y are listed. It turns out that determining the right capitalization of terms aids in filtering out incorrect links. WordNet synsets for 'house' will then only match articles about houses but not articles about movies or music singles called 'House'.

Cosine Similarity. The cosine $\mathbf{v}_x^T \mathbf{v}_y (||\mathbf{v}_x|| \; ||\mathbf{v}_y||)^{-1}$ between vectors $\mathbf{v}_x$, $\mathbf{v}_y$ derived for the short description gloss extracted from the English Wikipedia and the gloss/labels provided by WordNet, respectively. The vectors are created using TF-IDF scores after stemming using Porter's method.

Primary Sense Heuristic. The number of unqualified English Wikipedia labels where the WordNet synset is listed as the first (most frequent) noun sense in WordNet. A Wikipedia title like 'House' is considered unqualified if it does not include an additional qualification in parentheses, unlike 'House (novel)'. The most frequent sense of 'house' listed in WordNet is much more likely to correspond to Wikipedia's 'House' article than to pages with additional qualifications like 'House (1977 film)' or 'House (novel)'. The former reflects the most important meaning of a word as chosen by Wikipedia editors, and thus is more likely to correspond to the first sense listed in WordNet.

Together, these three signals allow us to learn whether a Wikipedia article and a WordNet synset describe the same thing.

3.1.4 *Redirect Matching*

Many projects treat redirects in Wikipedia as simple aliases for an entity. However, many redirects do not share the same referent with the page they redirect to. For instance, there are redirects from `Physisist` (i.e. human beings) to `Physics` (a branch of science) and from `God does not play dice` to `Albert Einstein`. There are large numbers of redirects from song names to album names or artist names, and so on. We decided to conservatively equate redirects with their targets only in the following two cases.

- the titles of redirect and redirect target match after parenthesized substring removal, Unicode NFKD normalization, diacritics and punctuation removal, and lower-case conversion
- the redirect uses certain templates or categories that explicitly indicate co-reference with the target (alternative names, abbreviations, etc.)

3.1.5 *Infobox Matching*

This function returns a constant $w > 0$ when an infobox template like `Infobox actor` is matched with an article or category having a corresponding title, in this case `Actor`, and 0.0 otherwise. We chose $w = 0.5$ because these links are not as reliable as interwiki links or redirect links. The function does not consider article titles with additional qualifications as matching, so `Actor (UML)` would not be considered.

3.2 Subclass Link Heuristics

Subclass linking functions use simple heuristics to link a class x to its potential parent classes y.

3.2.1 *Parent Categories*

This linker checks if categories are subclasses of their own parent categories as listed in Wikipedia. It first ensures that both x and y are likely to be categories denoting genuine classes. A genuine class like `Biologists` can have instances as its class members (individual biologists, ontologically speaking, are regarded as instances of `Biologists`). In contrast, other categories like

`Biology` or `Molecular biology` merely denote topic labels. It would be wrong to say that Charles Darwin "is a" `Biology`. For distinguishing the two cases automatically, we found that the following heuristic generalizes the singular/plural heuristic proposed for YAGO [39] to the multilingual case:

- headword nouns that are countable (can have a plural form) tend to indicate genuine classes
- headword nouns that are uncountable (exist only in singular form) tend to be topic tags

Hence, we take the titles of a category as well as its cross-lingual counterparts, remove parentheses, and rely on a dependency parser (if available) to keep only the headword. We then check that the headword is given in plural (for English), or is countable (in the general case). Countability information is extracted from WordNet and Wiktionary, the latter using regular expressions. We also added a small list of Wikipedia-specific exceptions (words like 'articles', 'stubs') that are excluded from consideration as classes.

3.2.2 Category-WordNet Subclass Relationships

If x is a category, then the headword of its title also provides a clue as to what parent classes are likely in WordNet. For instance, a category like `University book publishers` has 'publishers' as a headword. As earlier for the equality linker, we again relied on supervised learning to obtain a model that helps us disambiguate possible meanings of a word. This linker, too, relies on Ridge Regression [6] to learn likely meanings based on a labelled training set (see Section 5.3). As one of the main features, it uses

$$\sum_{l_x \in \mathrm{sim}(x)} \max_{l_y \in \mathrm{labels}(y)} w_x(l_x, x) w_y(l_y, y) \mathrm{sim}(l_x, l_y) \qquad (2)$$

This is similar to Equation 1, however $\mathrm{sim}(l_x)$ is used to obtain headwords of titles (with a dependency parser if possible). Additionally, $w_x(l_x, x)$ will be 1 if l_x is in plural or countable and 0 otherwise, allowing us to distinguish topic labels from genuine classes. A few exceptions are specified manually, e.g. 'physics', 'arts', and Wikipedia-specific ones like 'articles', 'templates'. $w_y(l_y, b)$ uses the number of alternative meanings as earlier. Apart from this, the linker also relies on the other features mentioned above for `equals`, e.g. cosine similarity.

3.2.3 WordNet Hypernymy

WordNet's notion of hypernymy between synsets is closely related to the `subclass` relation. This linking function hence returns 1 if y is a hypernym of x in WordNet, and 0 otherwise.

3.3 Instance Link Heuristics

Instance linking functions link individual entities to their classes.

3.3.1 Infoboxes

An "Actor" infobox in an article is a very strong indicator of the article being about an actor. The instance linker returns a constant $w_{\mathrm{infobox}} > 0$ if y is recognized as an infobox template that occurred on the page of the article associated with x, and 0 otherwise. Since infoboxes are just template invocations, heuristics need to be used to identify them. For this, we rely on a list of suffixes and prefixes (like "_Infobox") for different languages.

3.3.2 Categories

If y is a Wikipedia category for the article associated with x, this linking function assesses whether a headword of y (or of its interwiki translations) is in plural or countable, and returns 1 if this is the case, and 0 otherwise.

4. TAXONOMY INDUCTION

As shown in Figure 1, what we obtain using the linking functions is a large collection of statements connecting articles, categories, and infoboxes to each other and to WordNet by means of the `equals` relation (solid lines), statements connecting categories and WordNet entities to parent classes by means of `subclass`, and statements connecting articles to categories and infoboxes as `instance` links (both using dotted lines). However, due to the noisy heuristic nature of the connections and the fact that these entities come from different language editions of Wikipedia, it is not trivial to recognize that 'Fersental' is a valley rather than a language. In fact, in reality, we may have more than 100 languages and many more potential classes. What is needed is a procedure to aggregate information and produce the final, much more coherent knowledge base, which would ideally include the parts depicted in Figure 2. We proceed in two steps. The first step aggregates entity identifiers referring to the same entity by producing consistent equivalence classes. In the second step, taxonomical information from different linkers is aggregated to produce one single output taxonomy.

4.1 Consistency of Equality

In general, there will often be multiple entity identifiers that refer to the same entity and that are connected by `equals` statements. For instance, the German `Fersental` is equivalent to the corresponding Italian, Norwegian, and other articles about the valley. It will sometimes be convenient to jointly refer to the set of all of these equivalents. The symmetry and transitivity of equality leads us to the following definition to capture the connected components of the transitive closure of `equals`.

Definition 4. In a knowledge base $K = (\mathcal{U}, \mathcal{R}, \mathcal{S})$, an *e-component* $E \subseteq \mathcal{U}$ for some entity $x \in \mathcal{U}$ is a minimal set of entities containing x such that for all $x \in E$, $y \in \mathcal{U}$: statements $(x, r, y, w) \in \mathcal{S}$ or $(y, r, x, w) \in \mathcal{S}$ with $r =$ `equals`, $w > 0$ imply $y \in E$. We use the notation $E(x)$ to denote the e-component containing a node x.

Due to the heuristic nature of the equality linking functions, it often occurs that two entities x, y are transitively identified within an e-component, although it is known a priori that they should not be. For instance, we may have two different articles linked to the same WordNet synset. In some cases, the input from Wikipedia is imprecise, e.g. the English articles `Differential calculus` and `Derivative` are both linked to the same German-language article, despite being ontologically different.

Definition 5. The function $\delta(x, y)$ determines whether two entities x,$y \in \mathcal{U}$ should be separated. $\delta(x, y) = 1$ if and only if one of the following conditions hold (0 in all other cases).

a) x and y are two different WordNet entity identifiers
b) x and y are two different Wikipedia articles from the same edition of Wikipedia, and are not redirects of each other
c) x and y are two different Wikipedia categories from the same edition of Wikipedia, and are not redirects of each other
d) x has an interwiki link to a specific subsection within y, or vice versa
e) x is a Wikipedia disambiguation page and y is not recognized as one, or vice versa

Separating two such entities while removing a minimal (weighted) number of edges corresponds to computing minimal graph cuts. Unfortunately, we often have multiple pairs that simultaneously need to be separated. Computing a globally minimal cut corresponds to solving the Minimum Multicut problem, which is APX-hard and likely to be outside of APX [9]. To cope with this, we

first apply generic graph partitioning heuristics [13] to break up very large sparsely connected components into individual, much more densely connected clusters. On each of these densely connected clusters, we then apply a more accurate algorithm [12] with a logarithmic approximation guarantee. The implementation relies on CPLEX to obtain the fractional linear program solutions required by this algorithm (see our prior work [12] for details). In a few cases, the LP solver may time out, in which case we resort to computing minimal s-t cuts [15] between individual pairs of entities that should be separated. Minimal s-t cuts can be computed efficiently in $O(VE^2)$ or $O(V^2E)$ time. The statements corresponding to the cut edges are removed, and hence we obtain small e-components that should no longer conflate different concepts.

4.2 Aggregated Ranking

Having made the `equals` links consistent, we then proceed to build the class hierarchy. The algorithm generating the final output taxonomy should have the following three properties.

Property 1. The input links often link individual articles to their categories, but these categories might be language-specific local ones that are not part of a shared multilingual class hierarchy. The algorithm should hence consider not only immediate parents but also higher-order parents as candidate parents for the final output.

Property 2. Information derived from multiple sources should be given a higher weight than comparable information obtained only from a single source, as it is likely to be more reliable and more salient. For example, many Wikipedia editions describe the Colorado River as a river, but only few declare it to be a border of Arizona. The output should be a *ranked list* of parents with corresponding scores rather than a simple set.

Property 3. The output ranking needs to take into account that classes are not independent, but themselves can stand in a relationship, e.g. two different versions of Wikipedia may have what is essentially the same class (`equals` links) or classes that stand in a subclass relationship to each other (`subclass` links). For instance, information about a Malay article entity may benefit from information available about a corresponding English article entity, and vice versa. Also, if an article is found to be in the class `State capitals in Malaysia` in the English Wikipedia, then the possible parent class `State capital` from WordNet should also gain further credibility.

Taking these considerations into account requires going beyond conventional rank aggregation algorithms. We use a Markov chain approach that captures dependencies between parents.

Definition 6. Given a set of entities X and a target relation r (`subclass` or `instance`), the set of *parents* $P(X, r)$ is the set of all nodes x_m that are reachable from $x_0 \in X$ following paths of the form $(x_0, x_1, \ldots, x_m)$ with $(x_i, r_i, x_{i+1}, w_i) \in \mathcal{S}, w_i > 0$ for all $0 \leq i < m$, and specific r_i. The path length m may be 0 (i.e. the initial entity x_0 is considered part of the parent entity set), and may be limited for practical purposes. When producing subclass links as output ($r =$`subclass`), all r_i must be `subclass` or `equals`. When producing instance links as output ($r =$`instance`), the first r_i that is not `equals` must be an `instance` relation, and any subsequent r_i must be either `equals` or `subclass`.

Instead of operating on original sets of parent entities $P(X, r)$, we consider the corresponding set of *parent e-components* $\{E(x) \mid x \in P(X, r)\}$ (see Definition 4), which consists of the e-components for all $x \in P(X, r)$.

Definition 7. Given a root node x_0, a target relation r, and a corresponding set of parent e-components $\{E_0, \ldots, E_n\}$ (such that $x_0 \in E_0$), we define $w_{i,j} = \sum_{x \in E_i} \sum_{y \in E_j} \sum_{(x,r',y,w) \in \mathcal{S}} w$ for all i,j from 0 to n, where r' is `instance` if $i = 0$ and r=`instance`, and r' is `subclass` in all other cases (i.e. if $i > 0$ or $r =$`subclass`). We further define $\mathrm{out}(i)$ as $\{j \mid w_{i,j} > 0\}$.

Definition 8. Given an entity x_0, a target relation r, a corresponding set of parent e-components $\{E_0, \ldots, E_n\}$ ($x_0 \in E_0$), and a weight matrix $w_{i,j}$ characterizing the links between E_i, we define a Markov chain $(E_{i_0}, E_{i_1}, \ldots)$ as follows. The set $\{E_0, \ldots, E_n\}$ serves as a finite state space S, an initial state $E_{i_0} \in S$ is chosen arbitrarily, and the transition matrix Q is defined as:

$$Q_{i,j} = \begin{cases} \frac{w_{i,j}}{c + \sum_{k \in \mathrm{out}(i)} w_{i,k}} & j \neq 0 \\ \frac{c + w_{i,j}}{c + \sum_{k \in \mathrm{out}(i)} w_{i,k}} & j = 0 \end{cases}. \quad (3)$$

Note that root node's component E_0 is included as part of the chain. The probability mass received by E_0 rather than by genuine parents E_i ($i > 0$) reflects the extent of our uncertainty about the parents: For instance, if all immediate parents of the root node are linked with very low weights, then E_0 will attract a high probability mass. In the definition, c is the weight endowed to random restarts, i.e. transitions from arbitrary states back to E_0. Higher values of c lead to a bias towards more immediate parents of the E_0, while lower values work in favour of more general (and presumably more reliable) parents at a higher level. It is easy to see that the Markov chain is irreducible and aperiodic if $c > 0$, so a unique stationary distribution must exist in those cases.

THEOREM 1. *If $c > 0$, then the Markov chain possesses a unique stationary probability distribution π ($\pi = \pi Q$). A unique stationary probability does not necessarily exist if $c = 0$.*

PROOF. Since S contains only E_0 and the parents, S is finite and every state is reachable from E_0. Since $c > 0$, we obtain a non-zero random restart probability $Q_{i,0} > 0$ for every i, so from every state one can transition back E_0, and thus the chain is irreducible. Additionally, if $c > 0$, then the state E_0 is aperiodic (one can remain in E_0 for any amount of steps), and hence the entire chain is aperiodic. By the Fundamental Theorem of Markov chains, a unique stationary distribution exists. In contrast, if $c = 0$, then E_0 is likely to be a transient state, in which case a unique stationary distribution cannot exist. □

Therefore, we can use the stationary distribution of the Markov chain to rank parents of a root node with respect to their connectedness to the root node. Algorithm 4.1 captures the steps taken to induce the taxonomy. It begins by forming e-components, which become the entities of the output knowledge base. Entity identifiers can be chosen arbitrarily from within each component, or, more likely, one would prefer article titles in a specific language. Non-taxonomic statements like `means` statements that provide human-readable labels or statements capturing factual knowledge like birth dates of people are directly mapped to the e-components.

Then, for each e-component E, the heuristics described in Section 3.2 are used to assess whether E is likely to be a class. In accordance with the outcome of this assessment, the parents are retrieved and a Markov chain is constructed. The fixpoint of π is computed using the well-known power iteration method. After ranking, a pre-defined selection function $\sigma(\pi, r, E_0, \ldots, E_n)$ produces output statements of the form (E_0, r, E_i, w) for parent e-components E_i ($i > 0$) with r as either `instance` or `subclass`

Algorithm 4.1 Taxonomy induction

procedure TAXONOMY($\mathcal{U}_0, \mathcal{R}_0, \mathcal{S}_0, c, \sigma$)
  $\mathcal{U} \leftarrow$ consistent e-components formed from $\mathcal{U}_0$ and $\mathcal{S}_0$ ▷ as explained in Section 4.1
  $\mathcal{R} \leftarrow \mathcal{R}_0$ ▷ complete set of relations
  $\mathcal{R}_\mathrm{T} \leftarrow \{\texttt{equals}, \texttt{instance}, \texttt{subclass}\}$ ▷ set of taxonomic relations
  $\mathcal{S} \leftarrow \{(E, r, E', w) \mid (x, r, y, w) \in \mathcal{S}_0, x \in E, y \in E', r \notin \mathcal{R}_\mathrm{T}\}$ ▷ map all non-taxonomic statements
  for all E in $\mathcal{U}$ **do** ▷ for all e-components
    $r \leftarrow \begin{cases} \texttt{subclass} & \text{if } E \text{ likely to be a class} \\ \texttt{instance} & \text{otherwise} \end{cases}$ ▷ Heuristic from Section 3.2
    determine parent entities $P(E, r)$ of E ▷ as per Definition 6
    enumerate $\{E(x) \mid x \in P(E, r)\}$ as $E_0, \ldots, E_n$ ($E_0 = E$, other E_i arbitrary) ▷ the corresponding parent e-components
    construct Markov chain for E using $E_0, \ldots, E_n$ and c, r ▷ as per Definition 8
    choose arbitrary π such that $||\pi||_1 = 1, \pi_j \geq 0$ ▷ initial distribution
    repeat $\pi_0 \leftarrow \pi, \pi \leftarrow Q\pi$ **until** $||\pi - \pi_0||_2 < \epsilon$ ▷ Power iteration method (for some very small ϵ)
    $\mathcal{S} \leftarrow \mathcal{S} \cup \sigma(\pi, r, E_0, \ldots, E_n)$ ▷ choose suitable parents fulfilling some selection criterion
  return $K = (\mathcal{U}, \mathcal{R}, \mathcal{S})$ ▷ taxonomic knowledge base as output

and w derived from π. Usually, the top-ranked k e-components will be chosen, where $k = 1$ leads to a more traditional hierarchy, while higher k lead to more comprehensive knowledge bases. Filtering with respect to specific criteria can be performed, e.g. only classes with Chinese labels, or only WordNet synsets as classes, and of course filtering with respect to some minimal probability threshold. Although this process needs to be repeated for all e-components, this step is nevertheless not a bottleneck (see Section 5).

Properties of the Algorithm. The state space S includes not only immediate parents, but also e-components of superordinate parents (Definitions 6 and 8). For each suitable path from the root node x_0 to some superordinate parent x_m, all statement weights along the path are non-zero, so corresponding weights $w_{i,j}$ and state transition probabilities $Q_{i,j}$ must be non-zero. Hence the parent's state is reachable from the root node's state with non-zero probability, so Property 1 is fulfilled. Definition 7 implies that an e-component with similar input links from multiple children will have a higher $w_{i,j}$ than a comparable e-component with only one such incoming link, so Property 2 is satisfied. Finally, the aggregation into e-components accounts for `equals` dependencies between nodes, and the stationary distribution π of a Markov chain accounts for `subclass` dependencies ($\pi = \pi Q$), so Property 3 holds as well.

With these properties, the algorithm allows us to aggregate link information from heterogeneous sources, e.g. information from multiple editions of Wikipedia, including category and infobox information, and from WordNet. The output is a much more coherent taxonomic knowledge base.

5. EVALUATION

5.1 Dataset

We wrote a custom web crawler that downloads the latest Wikipedia XML dumps from Wikimedia's download site, retrieving 271 different editions of Wikipedia as of April 2010. The size of the uncompressed XML dumps amounts to around 89.55GB in total, out of which 25.4GB stem from the English edition.

5.2 Entity Equality

The linking functions provided 184.3 million directed interwiki links and 7.1 million other directed `equals` links. The WordNet disambiguation model was obtained by training on 200 manually labelled training examples, selected randomly among all Wikipedia articles and WordNet synsets sharing a label. The precision-recall curve on 207 random test examples (Fig. 3) shows the remarkably reliable results of the model, e.g. with a threshold of 0.5 we obtain 94.3% precision at 80.7% recall (F_1: 87.0%). The precision only drops sharply once we move towards recall levels significantly above 80%. The overall ROC AUC is 93.06%.

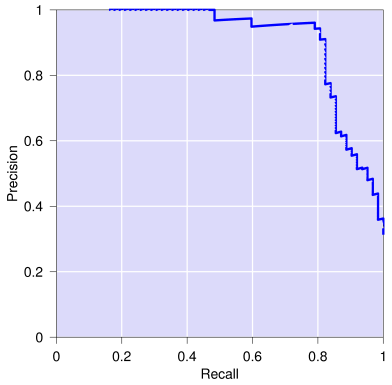

Figure 3: Precision-Recall curve for Wikipedia-WordNet links

The equality links led to 19.5 million initial e-components, including templates, categories, and redirects. It turns out that roughly 150,000 of these e-components contained nodes to be separated, among them a single large e-component consisting of nearly 1.9 million nodes. Overall, more than 5.0 million node pairs needed to be separated according to the criterion from Definition 5.

We applied the approach from Section 4.1 to separate the entities and obtain more consistent links. The process took several days to complete, with the expensive linear program solving with CPLEX (for the approximation algorithm) as the major bottleneck. We experimented with agglomerative clustering as an alternative, but found the solutions to be orders of magnitude worse in the number of edges to be cut. Using the approximation algorithm, a total of 2.3 million `equals` connections (4.6 million directed) were removed, resulting in 19.9 million e-components after separation.

5.3 Taxonomy

Linking Functions. As additional input to our ranking algorithm, the linking functions produced what correspond to 1.2 million `subclass` links and 20.1 million `instance` links between e-components. For the `instance` links, we chose $w_\mathrm{infobox} = 2$ because classes derived from infoboxes are more reliable than categories. The WordNet disambiguation for `subclass` was trained on 1,539 random mappings, the majority of these (1,353) being negative ex-

amples. On a test set of 234 random mappings, we obtain a precision of 81.3% at 40.0% recall, however going above 40% recall, the precision drops sharply, e.g. 60.8% precision at 47.7% recall. This task is apparently more difficult than the `equals` disambiguation, because less contextual information is directly available and because our heuristics for detecting classes may fail. Overall, there would be 6.1 million `subclass` links, but we applied a minimal threshold weight of 0.4 to filter out the very unreliable ones. The ROC AUC is only 65.8%. This shows that using these linking functions alone can lead to a taxonomy with many incorrect links.

Table 1: Ranked subclass examples

Class		WordNet parent	Wikipedia parent
Hamsters	1.	rodent	Rodents
	2.	hamster	Pets
	3.	mammal	Domesticated animals
Science museums in New Mexico	1.	museum	Museums
	2.	science museum	Science museum
	3.	depository	Museums in New Mexico

Table 2: Ranked instance examples

Entity		WordNet parent	Wikipedia parent
Fersental	1.	valley	Valleys
	2.	natural depression	Valleys of Italy
	3.	geological formation	Valleys of Trentino/ Alto Adige
The Spanish Tragedy	1.	book	Book
	2.	publication	British plays
	3.	piece of work	Plays

Table 3: Coverage of individual entities by source Wikipedia

	instances	WN instances	non-Engl. WN instances
English	3,109,029	3,004,137	N/A
German	911,287	882,425	361,717
French	868,864	833,626	268,693
Polish	626,798	579,702	159,505
Italian	614,524	594,403	161,922
Spanish	568,373	551,741	162,154
Japanese	544,084	519,153	241,534
Dutch	533,582	508,004	128,764
...	...	...	...
Total	13,982,432	13,405,345	2,917,999
E-components	5,790,490	5,379,832	2,375,695

Algorithm. We thus relied on our algorithm to choose reliable parents. In our experiments, the algorithm's c parameter was fixed at $c = \frac{1}{2}$, based on the intuition that if there is only one parent with weight 0.5, then that parent should be reached with probability $\frac{1}{2}$ from the current state. In order to increase the speed, we limited the maximal parent path length in Definition 6 to $m = 4$. This means that thousands of states that would obtain near-zero probabilities are pruned in advance. A second key to making it fast is relying on the fact that many entities share common parents, so the expensive parent lookups should be cached. This allowed us to process all 19.9 million e-components in less than 3 hours on a single 3GHz CPU. Examples of subclass and instance rankings are given in Tables 1 and 2, respectively, showing the highest-ranked WordNet and Wikipedia parents.

Coherence. Out of the 19.9 million e-components, a large majority consist of singleton redirects that were not connected to their redirect targets, due to our careful treatment of redirects in Section 3.1. For roughly 5.8 million e-components, we actually had `instance` links as input. To quantify the coherence, we determine what fraction of these e-components can be connected to e-components involving WordNet synsets, as WordNet can be considered a shared upper-level core. Table 3 shows that nearly all of these e-components are successfully attached to the shared upper ontology. The first column shows the number of entities for which we have `instance` links, while the second column is restricted to those for which we could establish `instance` links to WordNet (at a reachability probability threshold of 0.01). The small differences in counts between these two columns indicate that most entities for which there is any class information at all can be integrated into the upper-level backbone provided by WordNet. The third column lists the number of e-components that are independent of the English Wikipedia but nevertheless have successfully been integrated by our algorithm. While some fraction of those may correspond to entities for which cross-lingual interwiki links need to be added, large numbers are entities of local interest without any matching English Wikipedia article. Additionally, we found that 338,387 e-components were connected as subclasses of WordNet synsets, out of a total of 360,476 e-components with outgoing `subclass` links.

Table 4: Accuracy of Subclass Links To WordNet

top-k	Sample Size	Initial Links	Ranked Links
1	104	$82.46\% \pm 7.08\%$	$83.38\% \pm 6.92\%$
2	196	$57.51\% \pm 6.85\%$	$83.03\% \pm 5.17\%$
3	264	$45.89\% \pm 5.97\%$	$79.87\% \pm 4.78\%$

Accuracy. Table 4 shows a manual assessment of highest-ranked WordNet-based parent classes for over 100 random entities. We rely on Wilson score intervals to generalize our findings to the entire data set. For $k = 2, 3$, the ranked output is significantly more reliable than the $w_{i,j}$ between e-components resulting from the initial `subclass` links. The aggregation effect is even more noticeable for the `instance` links to WordNet in Table 5. To connect instances to WordNet, the algorithm needs to combine `instance` links with unreliable `subclass` links. Yet, the output is significantly more accurate than the input `subclass` links, for $k = 1$, 2, and 3. This means that the Markov chain succeeds at aggregating evidence across different potential parents to select the most reliable ones. We additionally asked speakers of 3 other languages to evaluate the top-ranked WordNet synset for at least 100 randomly selected entities covered in the respective language, but without corresponding English articles. We see that non-English entities are also connected to the shared upper-level ontology fairly reliably. The main sources for errors seem to be topic categories that are interpreted as classes and word sense disambiguation errors from the subclass linking function. Fortunately, we observed that additional manually specified exceptions as in YAGO [39] would lead to significant accuracy improvements with very little effort, as certain categories are very frequent.

Table 5: Accuracy of Instance Links To WordNet

Language	top-k	Sample Size	Wilson score interval
English	1	116	$90.05\% \pm 5.20\%$
English	2	229	$86.72\% \pm 4.31\%$
English	3	322	$85.91\% \pm 3.75\%$
Chinese	1	176	$90.59\% \pm 4.18\%$
German	1	168	$90.15\% \pm 4.36\%$
French	1	151	$92.30\% \pm 4.06\%$

Coverage. The total number of output e-components in MENTA is roughly 5.4 million excluding redirects (Table 3), so in terms of the number of entities and entity labels, this means that MENTA is significantly larger than existing multilingual and monolingual taxonomies relying only on the English Wikipedia, which as of June 2010 has around 3.3 million articles.

5.4 Upper-Level Ontology

Wikipedia as Upper Level. We can choose to retain WordNet as an integral core of MENTA, or alternatively, we may also create a more Wikipedia-centric version where WordNet only serves as background knowledge to help us connect different articles and categories and obtain a more coherent taxonomy. To achieve this, it suffices to have the selection function σ choose only e-components including Wikipedia articles or categories. This amounts to pruning all e-components that consist only of WordNet synsets without corresponding Wikipedia articles or categories. What we obtain is a taxonomy where the root node is based on the English article `Entity` and its equivalents in other languages. The upper level is shallower than with WordNet, as many different classes like `Organisms`, `Unit`, `Necessity`, are directly linked to `Entity`.

Lexical Knowledge. If we instead maintain all of WordNet at the top level, then after forming e-components, that part of our knowledge base can be considered a multilingual version of WordNet. A total of 42,041 WordNet synsets have been merged with corresponding Wikipedia articles or categories. We found that WordNet is extended with words and description glosses in 254 languages, although the coverage varies significantly between languages. The average number of Wikipedia-derived labels for these synsets is 20. In Table 6, the results are compared with UWN [11], a multilingual wordnet derived mainly from translation dictionaries. While MENTA's coverage is limited to nouns, we see that MENTA covers comparable numbers of distinct terms. The number of `means` statements is lower than for UWN, because each Wikipedia article is only merged with a single synset. The precision of MENTA's disambiguation is 94.3%, which is significantly higher than the 85-90% of UWN. This is not surprising, because an approach based on translation dictionaries has much less context information available for disambiguation, while MENTA can make use of Wikipedia's rich content and link structure. Additionally, MENTA's output is richer, because we add not only words but also have over 650,000 short description glosses in many different languages as well as hundreds of thousands of links to media files and Web sites as additional information for specific WordNet synsets. Gloss descriptions are not only useful for users but are also important for word sense disambiguation [25]. Finally, of course, our resource adds millions of additional instances in multiple languages, as explained earlier.

Alternative Upper-Level Ontologies. In an additional experiment, we studied replacing WordNet's lexically oriented upper-level ontology with the more axiomatic one provided by SUMO [29]. We added SUMO's class hierarchy as well as the publically available SUMO-WordNet mappings as inputs to the instance ranking, and found that SUMO can be extended with 3,036,146 instances if we accept those linked to a SUMO class with a Markov chain stationary probability of at least 0.01. The sampled accuracy of 177 top-1 links was $87.9\% \pm 4.7\%$. Problems often resulted from SUMO-WordNet mappings that did not reflect the meaning of a WordNet sense appropriately.

Table 6: Multlingual WordNet (upper-level part of MENTA)

Language	`means` Statements in MENTA	Distinct Terms in MENTA	Distinct Terms in UWN
Overall	845,210	837,627	822,212
French	36,093	35,699	33,423
Spanish	31,225	30,848	32,143
Portuguese	26,672	26,465	23,499
German	25,340	25,072	67,087
Russian	23,058	22,781	26,293
Dutch	22,921	22,687	30,154

5.5 Large-Scale Domain-Specific Extensions

A salient property of our approach is that we can easily tap on additional large-scale knowledge sources in order to obtain even larger knowledge bases. Of course, infobox attributes and other information as provided by DBpedia can easily be integrated. Additionally, we can rely on the many domain-specific knowledge bases in the Linked Data Web [7], which describe biomedical entities, geographical objects, books and publications, music releases, etc. In order to integrate them we merely need an `equals` linking function for all entities and `equals` or `subclass` links for a typically very small number of classes. Our entity aggregation from Section 4.1 will then ensure that the links are consistent, and our ranking algorithm from Section 4.2 will choose the most appropriate classes, taking into account the weights of the `subclass` links.

As a case study, we investigated a very simple integration of the LinkedMDB dataset, which describes movie-related entities. The `equals` links for instances were derived from the existing DBpedia links provided with the dataset, which are available for films and actors. Hence we only needed to specify two manual `equals` links for these two classes to allow all corresponding entities to be integrated. We obtain additional information on 18,531 films and 11,774 actors already in our knowledge base. Additionally, up to 78,636 new films and 48,383 new actors are added. Similar extensions of MENTA are possible for many other domains.

5.6 Entity Search

Knowledge bases like MENTA are useful for semantic search applications. For instance, the Bing web search engine integrates results from Freebase for queries like 'pablo picasso artwork'. In Table 7, we compare the numbers of instances obtained as results from the English Wikipedia with the numbers of instances in MENTA. The Wikipedia column lists the number of articles belonging to a given category in the English Wikipedia, while the MENTA columns list the number of instances in MENTA's aggregated ranking (with a minimum stationary probability of 0.01). Even if we consider only MENTA instances present in the English Wikipedia, we often find more instances than directly given there, because our approach is able to infer new parents for instances.

Table 7: Entity Search Query Examples

Query	Wikipedia	MENTA (English Wikipedia)	MENTA (All)
cities and towns in Italy	8,156	8,509	12,992
european newspapers	13	389	1,963
people	441,710	882,456	1,778,078
video games developed in Japan	832	775	1,706

6. RELATED WORK

A number of projects have imported basic information from Wikipedia, e.g. translations and categories [22, 36], or facts like birth dates in Freebase (*freebase.com*). Such resources lack the semantic integration of conflicting information as well as the taxonomical backbone that is the focus of our work. Apart from such facts, DBpedia [3] also provides an ontology, based on a set of manually specified mappings from infoboxes to a coarse-grained set of 260 classes. However, the majority of English articles do not have any such infobox information, and entirely non-English articles are simply ignored. DBpedia additionally includes class information from YAGO [39], a knowledge base that links entities from Wikipedia to an upper-level ontology provided by WordNet. We adopted this idea of using WordNet as background knowledge as well as some of the heuristics for creating instance and subclass links. YAGO's upper ontology is entirely monolingual, while in MENTA the class hierarchy itself is also multilingual and additionally accommodates entities that are found in non-English Wikipedias. Furthermore, the class information is simultaneously computed from multiple editions of Wikipedia. Nastase et al. [28] exploit categories not only to derive `isA` relationships, but also to uncover other types of relations, e.g. a category like 'Villages in Brandenburg' also reveals where a village is located.

There are other projects that have proposed heuristics for interlinking Wikipedia editions or linking Wikipedia to WordNet. Ponzetto et al. [34, 33] investigated heuristics and strategies to link Wikipedia categories to parent categories and to WordNet. Their results are very interesting, as they lead to a taxonomy of classes based on the English Wikipedia's category system, however they did not study how to integrate individual entities (articles) into this taxonomy. Wu and Weld [44] use parsing and machine learning to link infobox templates to WordNet. The Named Entity WordNet project [42] attempts to link entities from Wikipedia as instances of roughly 900 WordNet synsets. Others investigated heuristics to generate new cross-lingual links between different editions of Wikipedia [30, 38]. The focus in our paper is on a suitable technique to aggregate and rank information delivered by such heuristics, and many of these heuristics could in fact be used as additional inputs to our algorithm. The same holds for the large body of work on information extraction to find `isA` relationships in text [19]. Adar et al. [1] and Bouma et al. [8] studied how information from one Wikipedia's infoboxes can be propagated to another edition's articles, which is distinct from the problem we are tackling.

Concerning multilingual knowledge bases in general, previous results have been many orders of magnitude smaller in terms of the number of entities covered [24, 18], or lack an ontological class hierarchy [26]. EuroWordNet [43] and UWN [11] provide multilingual labels for many general words like 'university', but lack the millions of individual named entities (e.g. 'Napa Valley' or 'San Diego Zoo') that Wikipedia provides.

There are numerous studies on supervised learning of hierarchical classifications [14], but such approaches would require reliable training data for each of the several hundred thousand classes that we need to consider. Hierarchical agglomerative clustering has been used to derive monolingual taxonomies [23], however clustering techniques will often merge concepts based on semantic relatedness rather than specific ontological relationships. Our work instead capitalizes on the fact that reasonably clean upper ontologies already exist, so the main challenge is integrating the information into a coherent whole. Snow et al. [37] proposed a monolingual taxonomy induction approach that also considers the evidence of coordinate terms when disambiguating. Their approach assumes that evidence for any superordinate candidates is directly given as input, while our approach addresses the question of how to produce evidence for superordinate candidates based on evidence for subordinate candidates. For instance, very weak evidence that Stratford-upon-Avon is either a village or perhaps a city may suffice to infer that it is a populated place. Talukdar et al. [40] studied a random walk technique to propagate class labels from seed instances to other coordinate instances, but did not consider hierarchical dependencies between classes. Another interesting alternative approach, proposed by Wu and Weld [44], relies on Markov Logic Networks to jointly perform mappings between entities and derive a taxonomy. Unfortunately, such techniques do not scale to the millions of entities we deal with in our setting.

Our Markov chain algorithm is most similar to PageRank with personalized random jump vectors [31], however our transition matrix is based on statement weights, and the probability for returning to the root node depends on the weights of the alternative statements rather than being uniform for all nodes. Uniform weights mean that single parents are visited with very high probability even if they are only very weakly connected, while in our approach such irrelevant parents will not obtain a high transition probability. Other studies have relied on PageRank to find important vocabulary in an ontology [45] and to perform word sense disambiguation [2]. Our Markov chain model differs from these in that we look for salient parents for a specific node rather than generic random walk reachability probabilities. We are not aware of any Markov chain-based approaches for constructing class hierarchies.

7. CONCLUSIONS AND FUTURE WORK

We have presented techniques to integrate multilingual information into a single taxonomy. We succeeded in integrating 13.4 million out of 14.0 million possible articles from different Wikipedia editions into a single taxonomy. The result of this work is MENTA, presumably the largest multilingual lexical knowledge base, which is freely available for download at `http://www.mpii.de/yago-naga/menta/`.

In future work, we would like to investigate algorithms for an extended scenario where we assume that we are additionally given a set of class disjointness constraints (for example a human being cannot simultaneously be a geographical location) and need to post-process the rankings accordingly. Such extra information would allow us to improve the quality even further. We would also like to apply our approach to other types of input data, e.g. using large-scale information extraction techniques [41] to collect named entities and clues about their classes from text. Overall, we see our research as an important step towards new knowledge bases that integrate many existing large-scale data sources and offer more than the sum of the inputs.

8. REFERENCES

[1] E. Adar, M. Skinner, and D. S. Weld. Information arbitrage across multi-lingual Wikipedia. In *Proc. WSDM 2009*. ACM, New York, NY, USA, 2009.

[2] E. Agirre and A. Soroa. Using the multilingual central repository for graph-based word sense disambiguation. In *Proc. LREC 2008*, Marrakech, Morocco, 2008. ELRA.

[3] S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proc. ISWC/ASWC*, LNCS 4825. Springer, 2007.

[4] H. Bast, A. Chitea, F. Suchanek, and I. Weber. ESTER: Efficient search in text, entities, and relations. In *Proc. SIGIR*. ACM, New York, NY, USA, 2007.

[5] A. Bellaachia and G. Amor-Tijani. Enhanced query expansion in English-Arabic CLIR. In *Proc. DEXA 2008*, Washington, DC, USA, 2008. IEEE Computer Society.

[6] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1st ed. corr. 2nd printing edition, 2007.

[7] C. Bizer, T. Heath, and T. Berners-Lee. Linked data – the story so far. *Int. J. Sem. Web and Inform. Sys.*, 2009.

[8] G. Bouma, S. Duarte, and Z. Islam. Cross-lingual alignment and completion of Wikipedia templates. In *Proc. Workshop Cross Lingual Information Access*. ACL, 2009.

[9] S. Chawla, R. Krauthgamer, R. Kumar, Y. Rabani, and D. Sivakumar. On the hardness of approximating multicut and sparsest-cut. In *Proc. IEEE CCC*, 2005.

[10] D. Cook. MLSN: A multi-lingual semantic network. In *Proc. NLP*, 2008.

[11] G. de Melo and G. Weikum. Towards a universal wordnet by learning from combined evidence. In *Proc. CIKM 2009*, New York, NY, USA, 2009. ACM.

[12] G. de Melo and G. Weikum. Untangling the cross-lingual link structure of wikipedia. In *Proc. ACL 2010*, Uppsala, Sweden, 2010. ACL.

[13] I. S. Dhillon, Y. Guan, and B. Kulis. Weighted graph cuts without eigenvectors. a multilevel approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(11):1944–1957, 2007.

[14] S. T. Dumais and H. Chen. Hierarchical classification of Web content. In *Proc. SIGIR*, Athens, Greece, 2000. ACM.

[15] J. Edmonds and R. M. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM*, 19(2):248–264, 1972.

[16] O. Etzioni, K. Reiter, S. Soderland, and M. Sammer. Lexical translation with application to image search on the web. In *Proc. MT Summit XI*, 2007.

[17] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[18] C. Fellbaum and P. Vossen. Connecting the universal to the specific: Towards the global grid. In *Proc. IWIC*, volume 4568 of *LNCS*. Springer, 2007.

[19] N. Garera and D. Yarowsky. Minimally supervised multilingual taxonomy and translation lexicon induction. In *Proc. IJCNLP*, 2008.

[20] Z. Gong, C. W. Cheang, and L. H. U. Web query expansion by WordNet. In *Proc. DEXA 2005*, volume 3588 of *LNCS*. Springer, 2005.

[21] P. Hayes. RDF semantics. W3C recommendation, World Wide Web Consortium, 2004.

[22] D. Kinzler. Automatischer Aufbau eines multilingualen Thesaurus durch Extraktion semantischer und lexikalischer Relationen aus der Wikipedia. Master's thesis, Universität Leipzig, 2008.

[23] I. P. Klapaftis and S. Manandhar. Taxonomy learning using word sense induction. In *Proc. NAACL-HLT*. ACL, 2010.

[24] K. Knight and S. K. Luk. Building a large-scale knowledge base for machine translation. In *Proc. AAAI*, 1994.

[25] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proc. SIGDOC 1986*. ACM, 1986.

[26] Mausam, S. Soderland, O. Etzioni, D. Weld, M. Skinner, and J. Bilmes. Compiling a massive, multilingual dictionary via probabilistic inference. In *Proc. ACL-IJCNLP*. ACL, 2009.

[27] D. N. Milne, I. H. Witten, and D. M. Nichols. A knowledge-based search engine powered by Wikipedia. In *Proc. CIKM 2007*, New York, NY, USA, 2007. ACM.

[28] V. Nastase, M. Strube, B. Boerschinger, C. Zirn, and A. Elghafari. WikiNet: A very large scale multi-lingual concept network. In *Proc. LREC 2010*. ELRA, 2010.

[29] I. Niles and A. Pease. Towards a Standard Upper Ontology. In *Proc. FOIS*, 2001.

[30] J.-H. Oh, D. Kawahara, K. Uchimoto, J. Kazama, and K. Torisawa. Enriching multilingual language resources by discovering missing cross-language links in Wikipedia. In *Proc. WI/IAT*, Washington, DC, USA, 2008. IEEE.

[31] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.

[32] J. Pasternack and D. Roth. Learning better transliterations. In *Proc. CIKM 2009*, New York, NY, USA, 2009. ACM.

[33] S. P. Ponzetto and R. Navigli. Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proc. IJCAI*. Morgan Kaufmann, 2009.

[34] S. P. Ponzetto and M. Strube. WikiTaxonomy: A large scale knowledge resource. In *Proc. ECAI 2008*. IOS Press, 2008.

[35] N. Schlaefer, J. Ko, J. Betteridge, M. Pathak, E. Nyberg, and G. Sautter. Semantic extensions of the Ephyra QA system for TREC 2007. In *Proc. TREC 2007*. NIST, 2007.

[36] C. Silberer, W. Wentland, et al. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In *Proc. LREC*. ELRA, 2008.

[37] R. Snow, D. Jurafsky, and A. Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *Proc. ACL*, Morristown, NJ, USA, 2006. ACL.

[38] P. Sorg and P. Cimiano. Enriching the crosslingual link structure of Wikipedia. A classification-based approach. In *Proc. AAAI 2008 Workshop Wikipedia and AI*, 2008.

[39] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *Proc. WWW*. ACM, 2007.

[40] P. P. Talukdar, J. Reisinger, M. Paşca, D. Ravichandran, R. Bhagat, and F. Pereira. Weakly-supervised acquisition of labeled class instances using graph random walks. In *Proc. EMNLP 2008*, Morristown, NJ, USA, 2008. ACL.

[41] N. Tandon and G. de Melo. Information extraction from web-scale n-gram data. In *Proc. Web N-gram Workshop at ACM SIGIR 2010*.

[42] A. Toral, R. Muñoz, and M. Monachini. Named Entity WordNet. In *Proc. LREC*. ELRA, 2008.

[43] P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Springer, 1998.

[44] F. Wu and D. S. Weld. Automatically refining the Wikipedia infobox ontology. In *Proc. WWW*. ACM, 2008.

[45] X. Zhang, H. Li, and Y. Qu. Finding important vocabulary within ontology. In *Proc. ASWC 2006*, volume 4185 of *LNCS*. Springer, 2006.