

EXTRACTING MULTILINGUAL DICTIONARIES FOR THE TEACHING OF CS AND AI

Dimitar Kazakov

University of York

York, UK

YO10 5DD

kazakov@cs.york.ac.uk

<http://www.cs.york.ac.uk/~kazakov>

Ahmad R. Shahid

University of York

York, UK

YO10 5DD

ahmad@cs.york.ac.uk

<http://www.cs.york.ac.uk/~ahmad>

ABSTRACT

This paper describes a method for creating multilingual dictionaries using Wikipedia as a resource. A lucky strike on the road to multilingual information retrieval, the main idea is simple: taking the titles of Wikipedia pages in English and then finding the titles of the corresponding articles in other languages produces a multilingual dictionary in all those languages. While the page contents may vary greatly, the great majority of page titles seem to be faithfully translated. Here, the method is used to produce specialised dictionaries in two areas: Computer Science and Artificial Intelligence. Such dictionaries can form a useful auxiliary teaching resource by providing students from non-English speaking countries with a quick translation of some of the key concepts in the subject area.

Keywords

University education, Computer Science, Artificial Intelligence, multilingual dictionaries, English as a second language, Web crawler, Wikipedia, Web mining.

1. INTRODUCTION

We have previously demonstrated how multilingual dictionaries can be extracted from Wikipedia [5]. Taking the titles of Wikipedia pages in English and then finding the titles of the corresponding articles in other languages produces a multilingual dictionary in all those languages. A Web crawler – a specialised

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

© 2004 HE Academy for Information and Computer Sciences

program recursively traversing all referenced URLs in a hypertext document, is used to collect the data. Here, the method is used to produce specialised dictionaries in 36 different languages, and two areas: Computer Science and Artificial Intelligence.

2. LITERATURE REVIEW

There have been efforts in the past to build multilingual dictionaries with varying degrees of success, and the ones we know of are only extensions of bilingual dictionaries already available [2,3,4]. Yet none of them had tried to use Wikipedia as the potential source of lexical information. Only very recently, have there been attempts to tap into the multilingual aspect of Wikipedia, which has been used to identify named entities [1]. Richman and Schone proposed a method of using the multilingual nature of Wikipedia to annotate a large corpus of text with Named Entity Recognition (NER) tags in six different languages: French, Ukrainian, Spanish, Polish, Russian, and Portuguese [1]. They used the Wikipedia in those languages, and manually derived a small set of key phrases, identifying named entities. For instance, for the named entity "Person", the key phrases could be: "People by" or "Given names". For each article title of interest they extracted the categories to which that entry was assigned. For instance, any entry in "Category:Living People" will refer to a person, and it shows that the article referred to the named entity "Person". In case they failed to tag the entry based on initial categories found, they would go one level deeper, till either the suitable category is found or they hit the preset limit on how deep the search would go. For instance, in order to classify "Jacqueline Bhabha," the system extracted the categories "British Lawyers," "Jewish American Writers," and "Indian Jews." Yet none of them matched any of the key phrases, so it proceeded to extract second order categories "Lawyers by nationality," "British legal professionals,"

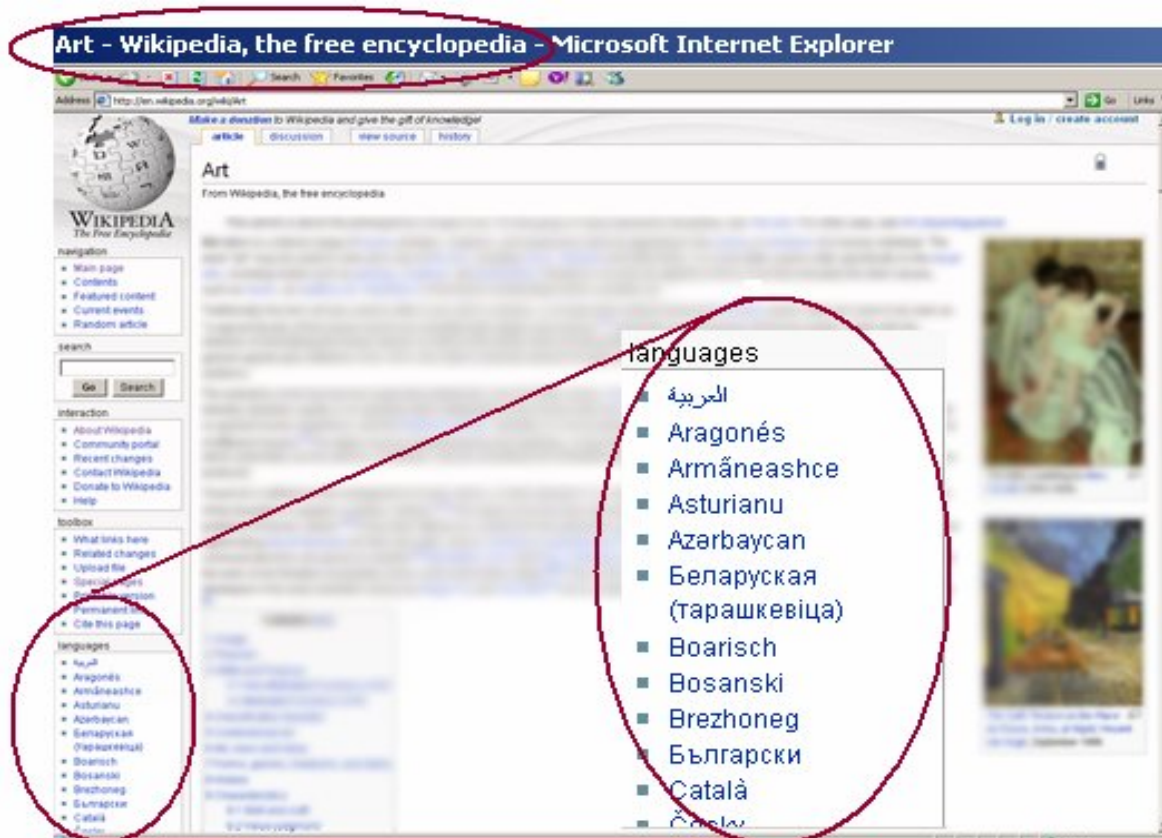


Figure 1: A Wikipedia Snapshot with links for pages in other languages.

“American writers by ethnicity,” “Indian people by religion,” etc. “People by” was on the list of key phrases and hence she was correctly classified. Finally they used Wiktionary, an online collaborative dictionary, to eliminate some common nouns. For multilingual categorization, they looked at the categories of the corresponding English titles. In case there was no English equivalent on Wikipedia, they would look at the non-English categories assigned to the title and try to determine the corresponding English categories and choose the most suitable category from amongst them. They demonstrated that Wikipedia based NER system performed comparably to one developed from 15-40.10³ words of human annotated newswire.

3. MULTILINGUAL LEXICON GENERATION

In this section, we give a brief account of our previous work, on which the approach used here is based.

3.1 Wikipedia

Since 2001, when it was created, Wikipedia has emerged as a huge online resource attracting over 684 million visitors yearly by 2008. There are more than 75,000 active contributors working on more than 10,000,000 articles in more than 250 languages (Wikipedia, August 3, 2008). Each

Wikipedia page has links to pages on the same topic in other languages, as shown in Figure 1. The figure also highlights the page title in the top left corner. The titles of all these pages can be extracted and combined in the form of tuples, which can be used as entries in the resulting multilingual dictionary, detailing a word in English and its translations in the other selected languages. In order to build the dictionary, a program (web crawler) was used.

3.2 Web Crawler

A web crawler is a computer program that follows links on web pages to automatically collect data (hypertext) off the internet. We use it here to move from one Wikipedia article to another, collecting the above mentioned tuples of word/phrase translations in the process.

Our version of the web crawler takes the starting page(s) as an input from the user. It visits the given page, and extracts all the links on that page and appends them to a list. Then it repeats the process for each link collected earlier, and visits them one by one, extracting the links and once again appending them to the list. Putting them at the end (e.g., making the list a queue) ensures that the search method adopted is Breadth First Search (BFS). In our context following BFS will explore a number of related concepts consecutively while Depth First Search (DFS) would drift-off any given topic. There

English	German	French	Polish	Bulgarian	Greek	Chinese
Wikipedia	Wikipedia	Wikipedia	Wikipedia	Уикипедия	Βικιπαίδεια	维基百科
Encyclopedia	Enzyklopädie	Encyclopédie	Encyklopedia	Енциклопедия	Εγκυκλοπαίδεια	百科全书
English language	Englische Sprache	Anglais	Język angielski	Аγγλίσки език	Αγγλική γλώσσα	英語
Venice	Venedig	Venise	Wenecja	Венеция	Βενετία	威尼斯
Film director	Régisseur	Réalisateur	Reżyser	Режисьор	Σκηνοθέτης	電影導演
Uniform Resource Locator	Uniform Resource Locator	Uniform Resource Locator	Uniform Resource Locator	Унифициран локатор на ресурси	Uniform Resource Locator	統一資源定位符
Web search engine	Suchmaschine	Moteur de recherche	Wyszukiwarka internetowa	Търсачка	Μηχανή αναζήτησης	搜索引擎
University	Hochschule	Université	Unwersytet	Университет	Πανεπιστήμιο	大學
Monopoly	Monopol	Monopole	Monopol	Монопол	Μονοπώλιο	壟斷
Computer	Computer	Ordinateur	Komputer	Компютър	Ηλεκτρονικός υπολογιστής	計算機
University of Oxford	University of Oxford	Université d'Oxford	Unwersytet Oksfordzki	Оксфордски университет	Πανεπιστήμιο της Οξφόρδης	牛津大学
Population density	Bevölkerungsdichte	Densité de population	Gęstość zaludnienia	Гъстота на населението	Πυκνότητα πληθυσμού	人口密度
Presidential system	Präsidentielles Regierungssystem	Régime présidentiel	System prezydencki	Президентска република	Προεδρική Δημοκρατία	總統制
Dictatorship	Diktatur	Dictature	Dyktatura	Диктатура	Δικτατορία	專政
European Community	Europäische Gemeinschaft	Communauté européenne	Wspólnota Europejska	Европейска общност	Ευρωπαϊκή Κοινότητα	歐洲共同體
Benazir Bhutto	Benazir Bhutto	Benazir Bhutto	Benazir Bhutto	Беназир Бхуто	Μπενάζιρ Μπούτο	贝娜齐尔·布托
Thomas Edison	Thomas Alva Edison	Thomas Edison	Thomas Alva Edison	Томас Едисън	Τόμας Έντισον	托马斯·爱迪生
Art	Kunst	Art	Sztuka	Изкуство	Τέχνη	艺术
California	Kalifornien	Californie	Kalifornia	Калифорния	Καλιφόρνια	加利福尼亚州
Buddhism	Buddhismus	Bouddhisme	Buddyzm	Будизъм	Βουδισμός	佛教

Figure 2: A Snapshot of the heptalex.

may be technical aspects related to the use of memory by each approach but we will not discuss them here.

The BFS approach was used in the following experiments. With the BFS capability thus incorporated, other lists were defined that would keep track of all the web pages that have already been visited thus keeping the code from revisiting them and extracting repeatedly the same tuples into the lexicon.

Apart from ensuring that no two entries were the same, the entries which did not convey any useful information had to be weeded out, as some of the links on the web pages are Wikipedia-specific and are not really useful for building the dictionary. Apart from these, no entry was entered into the lexicon if it was purely numeric.

3.3 Previous Results: a Heptalex

A general lexicon containing tuples corresponding to word/phrase translations in 7 different languages, English, German, French, Polish, Bulgarian, Greek and Chinese, was created. We call the result a *heptalex* [5]. Figure 2 shows a snap shot of the lexicon in a table. (UTF-8 was used as the coding scheme which makes possible writing characters in other non-English languages.)

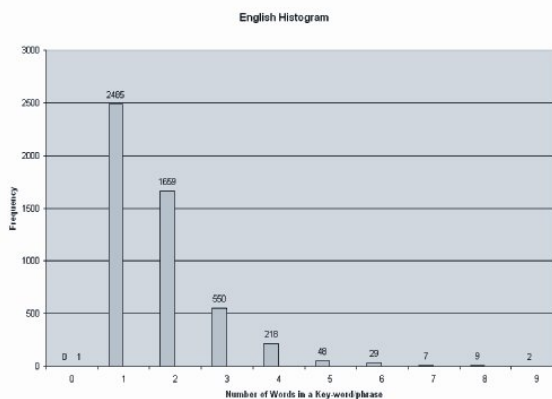


Figure 3: The English Histogram.

The goal was to extract as many entries as possible that provided translations of a concept in all seven languages. 5,006 unique entries were found as a result. In order to get that many entries, the crawler had to visit more than a quarter of the 2,000,000+ Wikipedia articles in English. The main limitation was Greek, which with some ~37,000 articles was the worst-represented language among the seven.

Using the least-prolific language as a pivotal language would have saved a lot of backtracking, yet English was chosen for its familiar alphabet, which also played a role in the hash tables used in the implementation.

Looking at Figure 3 one can see that unigrams make the bulk of entries (2,485 - almost 50% of the total), followed by bigrams (1,659 - 33% of the total).

In terms of semantics, the resulting lexicon is a mix of: toponyms, names of famous people, names of languages, and general concepts, such as “rock music” and “fire fighter”, among others.

4. EXTRACTING DOMAIN-SPECIFIC DICTIONARIES

We have already shown that building multilingual dictionaries from Wikipedia is feasible. With a general dictionary already created we decided to take it one step further by creating domain specific dictionaries. Such dictionaries can be useful for professionals of any particular domain; here we focus on the areas of Computer Science and Artificial Intelligence.

In order to create domain specific dictionaries the built-in *Category* attribute of Wikipedia was used. Wikipedia defines categories as tags to be used to ease navigation of pages. Even if one does not know that an article exists on any particular topic one can look into categories to look for interesting information. A programmer interested in *Functional programming* might look into this category and then look for subcategories of interest.

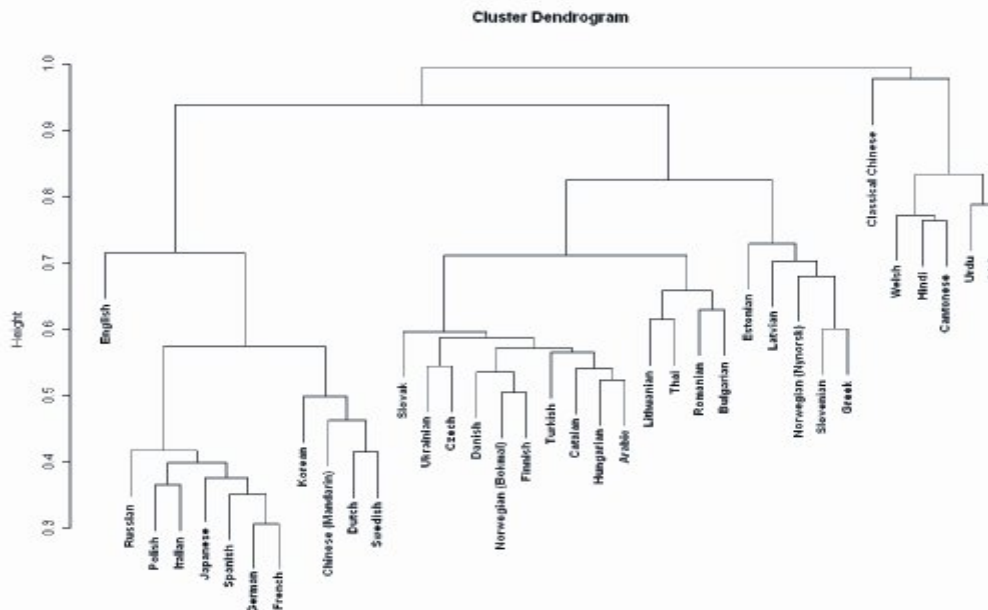


Figure 4: The Language Clusters for the CS domain.

4.1 A Computer Science Dictionary

Our first effort was to extract a dictionary of Computer Science (CS) related terms, by, narrowing down the list of categories which we were interested in. It was done manually by short listing categories from a complete list of categories provided on Wikipedia for reference, e.g., *.NET* and *Alan Turing* to show but two examples of the categories chosen. The result is a 37-language, 2,500 entry dictionary available from our Web pages.

4.2 Category Translations for CS & AI

Wikipedia assigns one or more categories to each article depending on which general category it might belong to. For instance, the article on Heuristic algorithm belongs to two different categories: Algorithms and Heuristics. That information is at times much more useful than merely looking at all the articles that might come up at any time while crawling. The categories information is organized and hierarchical and is thus quite useful in building dictionaries such as this one where we are looking at the subcategories of different categories.

Wikipedia also builds hierarchical structures bringing categories in one particular domain under one banner. Thus one can find the relevant subfields of Computer Science, such as Artificial Intelligence and Algorithms on the Computer Science category page. These subcategories in turn define further

subcategories taking it as many levels deep as the contributors of Wikipedia might like it to be.

Here we have looked at two domains: Computer Science (CS) and Artificial Intelligence (AI). We

have taken translations of categories and subcategories and combined them in the form of a lexicon. Once again nulls were allowed so as not to end up with very few entries.

CS is much more general than AI and hence its subcategories go many levels deep. AI on the other hand, being itself a subcategory of CS, does not go as deep. Thus for the former we just considered the categories, while for the latter we considered not just the categories but also the leaf nodes (individual Wikipedia articles), but only on the first page of the AI category.

Here 36 different languages have been catered to, removing Classical Chinese because of very few entries in the previous experiment.

More than 2,000 CS related categories were extracted in 36 different languages, which once again cover the wide spectrum of languages and scripts, both oriental and occidental. Interestingly, Chinese was found to be better represented language than German in this area, probably due to the inherent bias in the domain chosen.

There were very few translations found for any language. French, which otherwise had the second highest number of entries, second only to English, had just 21.3% of the translations. That shows how sparse the information is in these languages in the CS domain.

For AI around 450 categories were explored before the program started wandering off to other domains. That is far less than that for CS, as expected. Thus leaf nodes were also considered in this case. As percentage of English entries in this domain, Japanese came out at top but with mere 16%

entries. Japanese is expected to be well represented in this domain for the amount of work they have done. Yet a figure of 16% is quite low.

Snapshots of the selection of entries in CS and AI are shown in figures 7 and 8. Both dictionaries are freely available from the second author's Web site.

5. DISCUSSION

There are two issues we want to discuss here – how we can measure the potential of such an approach to producing dictionaries for any pair of languages, and how relevant such dictionaries may be to teaching in academia.

5.1 Measures of Usefulness

Asking oneself how useful this approach will be for any two languages in Wikipedia, one can think of two factors: (1) how well represented each of the languages is, i.e., how many entries there are in it, and (2) what is the likelihood that a concept, present in the first language, will also be present in the second. To elaborate on that second point, two different pairs of languages with the same total number of entries may show a different degree of overlap, which we define here as the ratio of the number of entries present in both languages to the number of entries (topics) that exist for at least one of the two languages:

$$\frac{|L_1 \cap L_2|}{|L_1 \cup L_2|}$$

One could expect that two languages are more likely to overlap in their Wikipedia articles, if they share a

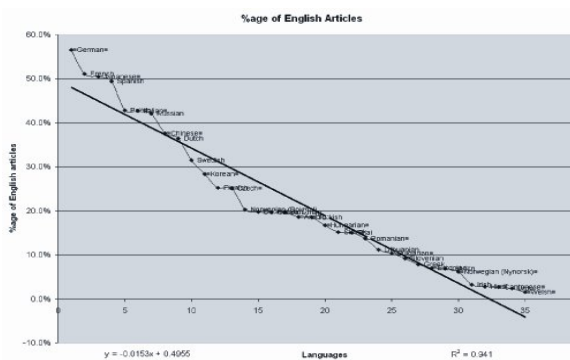


Figure 5: Ratio of articles for each language w.r.t. English common cultural background - for instance, both Latin and Italian are likely to include several pages on the Roman Catholic Church.

The languages that are linguistically related (and therefore easier to learn) may also show an overlap due to the fact that articles in one language are more easily translated in the other.

As an approximation, we can measure this similarity (min=0, max=1) on a sample of the Wikipedia pages. Using the ~2,500 entries of the CS dictionary described in Section 4.1, we produced a clustering tree as shown in Figure 4. There may be some

relationship between geographical distance - French and German - and overlap, as well as linguistic relatedness, Slovak-Ukrainian-Czech, Turkish-Finnish-Hungarian – but this is far from clear.

Two different graphs were plotted: One depicting the number of entries for each language in the lexicon as percentage of the English entries (Figure 5); and the other depicting the total number of articles in each language on Wikipedia (Figure 6). In each of the two graphs the two outliers (English and Classical Chinese) were removed and trend lines were plotted.

Figure 5 shows that the expected overlap between English and another language is a decreasing linear function of that language's rank, even if the number of articles as related to the rank (Figure 6) is a nonlinear function. In Figure 6, numbers represent languages. For instance, 2 represents German and 36 represents Irish.

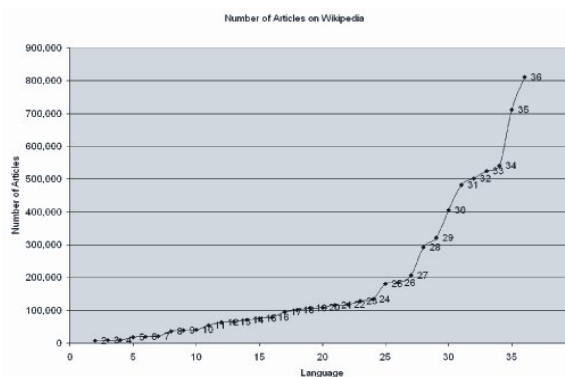


Figure 6: Number of articles in Wikipedia.

5.2 Relevance to Teaching

Such dictionaries can form a useful auxiliary teaching resource by providing students from non-English speaking countries with a quick translation of some of the key concepts in the subject area. Unlike Wikipedia, from which it was extracted, this data can be printed out and carried to lectures and labs. It provides three-fold assistance – firstly, it is often the case that the students know the concept in their own language, but do not know its English translation, a situation that is especially common among first-year undergraduates. Even when the English term is known, it may be easier to memorise it if it can be linked to items studied in one's mother tongue. Finally, it is not uncommon for someone who has studied abroad to return to their home country only to discover that specialised discussions in their native language are quite difficult, as they struggle to translate the concepts they have only encountered in the language in which they have received their degree. These dictionaries should help the students in all three aspects, and perhaps even encourage some of them to supply some of the missing entries in Wikipedia as they become better acquainted with the discipline studied.

6. REFERENCES

- [1] Alexander E.R. and Patrick S., Mining Wiki Resources for Multilingual Named Entity Recognition, *Proceedings of ACL-08: HLT*. (2008).
- [2] Christian B., Mathieu M. and Gilles S., The PAPHON Project: Cooperatively Building a Multilingual Lexical Database to Derive Open Source Dictionaries & Lexicons, *Proceedings of the 2nd Workshop NLPXML 2002, Post COLING 2002 Workshop*. (2002).
- [3] James B., JMdict: a Japanese-Multilingual Dictionary, *Coling 2004 Workshop on Multilingual Linguistic Resources*. (2004).
- [4] Mathieu L., Multilingual Dictionary Construction and Services Case Study with the Fe* Projects, *Proceedings of PAFLING'97*. (1997).
- [5] Shahid, A. and Kazakov, D., Automatic Multilingual Lexicon Generation using Wikipedia as a Resource, *in manuscript*.

English	German	Japanese	Chinese	Arabic	Korean	Bulgarian	Thai	Greek
Computer science	Informatik	計算機科学	计算机科学	حوسبة	전산학	Информатика	วิทยาศาสตร์คอมพิวเตอร์	Επιστήμη υπολογιστών
Computer architecture	Rechnerarchitektur	コンピュータアーキテクチャ	電腦架構	null	컴퓨터 구조	null	null	null
Semantics	Semantik	意味論	null	null	의미론	Семантика	null	null
Algorithms	Algorithmus	アルゴリズム	算法	خوارزميات	알고리즘	Алгоритми	ขั้นตอนวิธี	Αλγόριθμοι
Artificial intelligence	Künstliche Intelligenz	人工知能	人工智能	ذكاء اصطناعي	인공지능	null	ปัญญาประดิษฐ์	Τεχνητή νοημοσύνη
Computer programming	Programmierung	プログラミング	程序设计	برمجة	컴퓨터 프로그래밍	null	การเขียนโปรแกรม	null
Operating systems	Betriebssystem	オペレーティングシステム	操作系统	نظم تشغيل	운영 체제	null	ระบบปฏิบัติการ	null
Programming languages	Programmiersprache	プログラミング言語	程序设计语言	اللغات برمجة	프로그래밍 언어	Езици за програмиране	ภาษาโปรแกรม	Γλώσσες προγραμματισμού
Linux	Linux	Linux	Linux	لينكس	리눅스	ГНУ/Линукс	null	null
Cryptography	Kryptologie	暗号技術	密码学	تشفير	암호학	Криптография	null	Κρυπτογραφία

Figure 7: Selection of categories for Computer Science.

English	German	French	Japanese	Chinese	Arabic	Korean	Bulgarian	Thai	Greek
Artificial intelligence	Künstliche Intelligenz	Intelligence artificielle	人工知能	人工智能	ذكاء اصطناعي	인공지능	Изкуствен интелект	ปัญญาประดิษฐ์	Τεχνητή νοημοσύνη
Chess	Schach	Échecs	チェス	国际象棋	شطرنج	체스	Шахмат	null	Σκάκι
Game theory	Spieltheorie	Théorie des jeux	ゲーム理論	博弈论	نظرية الألعاب	게임 이론	Теория на игрите	ทฤษฎีเกม	Θεωρία παιγνίων
Search algorithms	null	Algorithme de recherche	検索アルゴリズム	搜尋演算法	null	검색 알고리즘	null	null	null
Machine learning	Maschinelles Lernen	null	機械学習	机器学习	تعلم آلي	기계 학습	null	การเฝ้าเรียนรู้ของเครื่อง	null
Robotics	Robotik	Robotique	ロボット工学	机器人学	علم الآلي	로봇공학	Роботика	null	null
Computer vision	Maschinelles Sehen	Vision par ordinateur	コンピュータビジョン	计算机视觉	رؤية حاسوبية	null	null	คอมพิวเตอร์วิทัศน์	null
Fuzzy logic	Fuzzy-Logik	Logique floue	ファジイ論理	模糊逻辑	منطق ضبابي	퍼지	null	ตรรกศาสตร์คลุมเครือ	null
Taxonomy	Taxonomie	Taxinomie	分類学	生物分類學	null	분류학	Таксономия	อนุกรมวิธาน	Ταξινόμια

Figure 8: Selection of categories for Artificial Intelligence.