



Eötvös Loránd University
Faculty of Informatics
Department of Software Technology
and Methodology

Intelligent system for catalyzing and disseminating collaborative work of heterogeneous expert communities

Péter Rabi

first-year Software Information Technologist M.Sc.

Anita Verő

first-year Software Information Technologist M.Sc.

Supervisors

András Lőrincz
Senior Research Fellow

Balázs Pintér
Ph.D. Student

Budapest, January 9, 2013

Abstract

Reliable and up-to-date information is essential in many fields. Particularly, in the domain of robotic surgery, there is an increasing need for such a knowledge base. We consider the problem of building an accurate and evolving knowledge repository from the collective intelligence of experts. We have developed (a) a system that brings together an expert community, who edit the data collaboratively; (b) a methodology and a portal system for the management of a community's collective intelligence with the support of social computing and Natural Language Processing (NLP) methods. Reliability is achieved by distilling the collective intelligence into scholarly papers through voting by the experts, selection by an Editorial Board, and a high quality peer-review process. Our community aims self-strengthening by recruiting experts from Wikipedia and spreading domain information to experts of different fields: terminological terms are explained by machine learning algorithms. We show how the system works in practice by analyzing the initial activities of the community in different parts of the system upon launching.

Contents

1. Tézisek	3
2 Introduction	5
3 Methods	7
3.1 Supporting the community	7
3.2 Implementation	8
4 Results	10
4.1 Components	11
4.1.1 Knowledge Representation	11
4.1.2 Community	12
4.1.3 Artificial Intelligence	13
4.2 Interactions	15
4.2.1 Interaction between the community and the knowledge representation	15
4.2.2 Interaction between the community and the AI	18
4.2.3 Interaction between the AI and the knowledge repre- sentation	20
4.3 Initial activities on the EuroSurge Portal	20
5 Discussion and planned work	21
6 Conclusion	22
References	24
A Word Explanation explained	24
A.1 The Method	25
A.2 Compiling the dictionary	27
B System internals	28

1. Tézisek

Dolgozatunknak két fő célja van: (1) dinamikusan fejlődő, megbízható tartalommal bíró tudásbázist hozunk létre és azt folyamatosan aktualizáljuk, (2) a működtetésben aktív kooperáló közösség munkáját segítjük természetesnyelv-feldolgozási eszközökkel. Kidolgozott módszertanunkat ismertetjük a dolgozatban. A módszereket megvalósító, portált is magában foglaló rendszer fontosabb elemeit létrehoztuk és bemutatjuk. A portál elemei mellett a közösséget intelligens eszközök támogatják, amelyek – elképzeléseink szerint – segítik, hogy az együttes tudás beépülhessen a munka során. Az együttes tudás megjelenésének segítése mellett egyéb tudás integrálását is segítik az intelligens eszközök kétféle reprezentáció egyidejű használatával. Ezek: szemantikus információ (ontológia), valamint összekapcsolt szöveges információ (wiki). A szöveges forma lesz a letisztult tudást tartalmazó kiterjesztett/egyesített ontológiák mögötti közös nevező. Az elgondolás hátterét Pintér és munkatársai dolgozták ki [7].

Rendszerünk egyesíti a Wikipedia és a Scholarpedia előnyeit a tartalom létrehozásában, a minőség javításában, és a végleges tartalom anonim peer review formában történő minőségellenőrzésében. Sikeres közösség elektronikusan és a közösség által (is) szerkesztett folyóiratot hozhat létre, amely (i) együttesen szerkesztett, beküldött, vagy meghívásos tudományos publikációkat tartalmazhat és (ii) intelligens eszközeink segítségével az érdeklődő felhasználó számára a tartalmat személyre szabható módon magyarázhatja. A publikációra érett tartalom megjelölése és a megírandó publikáció szerzőinek kiválasztása szavazáson alapul.

A dolgozatban bemutatásra kerülő eredmények és megvalósított funkciók a EuRoSurge EU FP7 CSA projekt keretein belül készültek.

Munkánkkal kapcsolatban az alábbi téziseket fogalmazzuk meg:

- *Együttes munka támogatása:*

A szakértők közötti egyeztetést a gépi tudást reprezentáló ontológiából és wikiből kinyert magyarázatokkal segíthetjük. A szakértők a magyarázatokat minősíthetik, ami az ellentmondó értelmezésekre és a gépi tudás hiányosságaira derít fényt. Ennek eredményeként, mind az ontológia, mind a szöveges tartalom javulhat. Az egyeztetés működését a Wikipedia, a DBPedia és a EuRoSurge portál tartalma illesztésének már megvalósult lépésein keresztül mutatjuk be.

- *Minőségellenőrzés módszerei:*

A szöveg alapú reprezentáció minőségellenőrzését a szavazórendszer és szerkesztőbizottság valósítja meg. A tartalomszerkesztők munkájának végső ellenőrzése a szerkesztőbizottság feladata. Bemutatjuk az első tartalom fejlődését. Listázzuk az európai projektben vállalt további (minimális) tartalmakat.

A mi munkánk

1. az ontológia kiterjesztését támogató félautomata eszközök, azaz automatikus kontextus kinyerést végző szoftver
2. az intelligens funkciókat segítő adatgyűjtési szoftver, aminek segítségével a Wikipedia 40 GByte-nyi tartalmából többek között az ismeretlen szavak értelmezését lehetővé tevő adatbázist készítettük: a szavak azonos értelmeihez tartozó kontextusokat összegyűjtöttük.
3. a közösség (szerkesztőbizottság, szerzők, tartalomszerkesztők, olvasók) munkafolyamatainak minőségellenőrző funkcióit támogató rendszer kidolgozása
4. a közösségi munka kezdeti lépéseinek dokumentálása és bemutatása.

2 Introduction

We need current information in every field of life. This is especially true in the domain of robot assisted surgery. One lab’s innovation in robotic technology relies on the results of others. A new surgical method invented in London can save lives in Berlin. Easily accessible and reliable information can be the motor of business between firms spread all across Europe. There is a need for a system for catalyzing and disseminating collected work of heterogeneous expert communities.

A static knowledge base is not a sufficient solution. The costs of professionals building it is very high. After running out of money (this usually happens when the project that funded the experts is over), the cumulated information quickly becomes obsolete. To overcome these disadvantages, a dynamically developing, cost-efficient, and sustainable solution should be considered. A collaboratively edited resource meets these requirements.

The driving force of such a knowledge base is a self-sustained community. Bringing experts together and keeping the online society live need carefully designed incentives [2]. Experts can be motivated by providing opportunity to increase social capital and reputation, while firms with interest in robotics can be attracted by a trusted marketing platform.

The quality of the knowledge base is of paramount importance because mistakes made during a surgical intervention can cost human lives. To secure high quality, we turn to methods employed by scientific journals. An editorial board – constituted of the top experts of the field – together with peer-review can ensure the high quality of contributions by weeding out low quality contributions. To lessen the work of reviewers and members of the editorial board, we utilize the collective intelligence of experts: they decide by vote whether a collaboratively edited *draft* has advanced enough to become a peer-reviewed *article*, and who the author of the article will be. Editorial board members consider these suggestions, but – as quality is our primary concern – they always have the last word.

Besides experts, we also consider non-experts who would like to access and understand information in the knowledge base. As robotic surgery is a multidisciplinary field, even practitioners can be non-experts. For example, surgeons may use terminological terms not understood by robot engineers and patients. Machine learning algorithms can help overcome this problem by explaining unknown terms with related concepts and their short description from the knowledge base. This functionality can accelerate the informa-

tion exchange between different domains and help disseminate the collected information in the non-expert community.

The knowledge base consists of texts in the form of semantic wiki documents and a formal ontology embedded into these texts. The ontology provides an unambiguous, formal definition of concepts in robotic surgery. The texts define the concepts in natural language. Both are collaboratively edited and used by the community and the artificial intelligence.

In this paper, we propose a methodology to distill a heterogeneous community’s collective intelligence into verified knowledge of high quality, and implement it in portal system. To manage the crowdsourcing process, we use a wiki: a web-based, collaborative content authoring system [4]. We merge the basic principles of two widely used wikis: Wikipedia¹ and Scholarpedia². Thus, we create an ecosystem to enable free, collaborative development of the knowledge base. The community can harness the advantages of both explicit and implicit coordination [5], similarly to Wikipedia.

Inspired by Scholarpedia, quality control instruments are responsible for filtering the collected knowledge. The result is in-depth, reliable content. We apply incentives to bring experts together and keep the society live. Intelligent features employ algorithms from machine learning and natural language processing to support everyday tasks and communication between members.

The first portal that supports this methodology has been implemented in the European Robotic Surgery (EuroSurge) EU FP7 project. Its goals coincide with the goals of the project: to gather robotic surgery experts in Europe, and create a mutual understanding of key concepts in the form of the knowledge base.

At this stage, the portal is being used by a small community: members of the project’s consortium and invited experts in the field of robotic surgery. The initial activities and experience of this community support the view that the methodology is viable and helpful in achieving our goals (Section 4).

An additional portal is planned in the domain of educational games. Here, we will support a community of education experts and developers in helping children develop their skills.

The content of this paper is organized as follows. In Section 3, we describe the considerations that led to the proposed system from the aspect of

¹http://en.wikipedia.org/wiki/Main_Page

²<http://scholarpedia.org>

the community and of the implementation. The results are presented in Section 4: we specify our methodology’s three main components, describe the features of our system in terms of their interactions, and analyze the initial activities of the community. We discuss the results and possible future improvements in Section 5. We conclude our work in Section 6. Finally, we give details about the implementation of the portal in Appendix B, and about the mathematical background of the “word explanation” feature in Appendix A.

3 Methods

In this Section, we examine our goal – to distill the collective intelligence of a heterogeneous community into verified knowledge of high quality – from two viewpoints: usability (how to support the community) and implementation (how to implement the methodology into a portal).

3.1 Supporting the community

To facilitate ease of use and reach a wide audience, we decided to base our system on the most widely known and used tool for collaborative editing: a Wiki [4].

The quality of the knowledge representation is of paramount importance because mistakes made during a surgical intervention can cost human lives. To secure high quality, we turn to methods employed by scientific journals and Scholarpedia³.

We introduce two different levels of information to our system: drafts and articles. Drafts allow easy contribution by a wide range of contributors: they are freely edited by anyone who has login access to the system. In contrast, articles are written by top experts in the field, and undergo a peer-review process before publication. Each article starts as a draft; as a draft matures, it may become an article.

We help the community make good decisions in matters that affect quality – which drafts become articles, who the authors of these articles will be, etc. – in a similar, two stage process. In the first stage, anyone who contributed can influence in a decision through a voting system. In second stage, an editorial board takes into account these suggestions, and decides. This design ensures

³<http://scholarpedia.org>

both that everyone participates in a decision and erroneous decisions are caught by the editorial board.

A primary concern in a multidisciplinary and highly technical field such as robotic surgery is communication. Communication is impeded between experts and non-experts, but even between experts (e.g., between a surgeon and a roboticist). Natural language processing can help overcome this problem by explaining unknown terms with related concepts and their short description from the knowledge base. This functionality can accelerate the information exchange between different domains and help disseminate the collected information in the non-expert community.

3.2 Implementation

The first portal that supports this methodology has been implemented in the European Robotic Surgery (EuroSurge) EU FP7 project. Its goals are the same as the goals of the project: to gather robotic surgery experts in Europe, and create a mutual understanding of key concepts in the form of the knowledge base.

Because of its potentially huge user base, the portal needs to be scalable and it needs to contain robust, industry-standard technologies. We chose to build our portal by integrating the Wordpress and the MediaWiki frameworks. There is also a third component called the “backend” that contains all computationally intensive features. The architecture is illustrated on Figure 1. Table 1 contains the enumeration of external software that the system relies on.

The workflow that assures quality is implemented as a MediaWiki extension. It consists of different techniques for writing an extension⁴, provided by the MediaWiki framework. It includes Hooks for injecting custom php code at key points within MediaWiki processing and Ajax to let JavaScript code interact with server side methods. It creates a Special Page subclass that provides a wiki page with content generated dynamically from the current state (by querying the database). Besides MediaWiki tables, it uses its own data tables to store operational data and logs in the MySQL database of MediaWiki.

The backend is implemented in Python and contains a wide variety of

⁴http://www.mediawiki.org/wiki/Manual:Developing_extensions#Extension_types

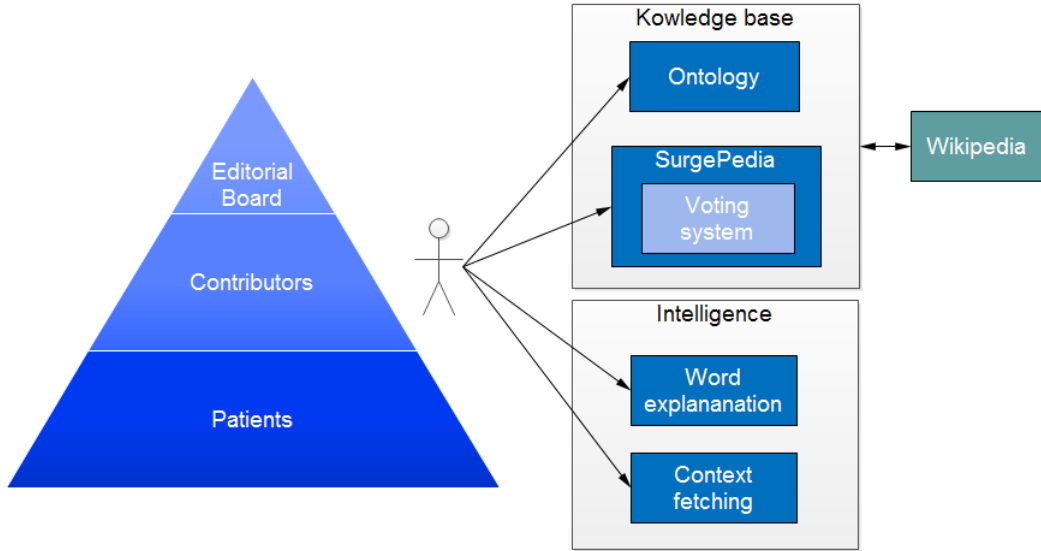


Figure 1: The architecture of the EuRoSurge portal. On the left, we present the hierarchy of users. On the right, every blue box represents a component of the system. The knowledge base consists of the ontology and a Wiki based textual representation. Users interact with the latter through SurgePedia, which itself consists of different sub-components, like the voting system. The backend contains the computationally intensive features: the word explanation and the context fetcher.

components, from information retrieval to numerical algorithms such as group Lasso. The component that processes Wikipedia and the component to help the extension of the ontology are built upon the widely used Lucene and SOLR information retrieval frameworks.

The sub-system facilitating the definition of new ontology concepts consists of two components: a SOLR⁵ deployment and a Python script. The SOLR instance is configured to index English PDF documents received in HTTP PUT messages. It is queried with expressions arriving as HTTP GET messages, for which it provides contexts in plain text format. The whole indexing and fetching process is managed by the Python script.

The word explanation feature is implemented as a web service in Python. It is based on web.py⁶, the robust web framework also used by reddit.com.

⁵<http://lucene.apache.org/solr/>

⁶<http://webpy.org/>

Framework	Examples
MediaWiki	Wikipedia
WordPress	Reuters, CNN
web.py	reddit.com
Lucene, SOLR	eBay
CAS	Moodle
MySQL	Industry standard

Table 1: Open source software frameworks integrated into our system. The left column enumerates the software; the right column gives examples of leading web portals built with these components.

The operation of the word explanation feature can be divided into two parts: offline and online. Offline operation prepares the necessary data on which the online operation relies. This is a mathematically involved method based on computational linguistics resources fetched from Wikipedia. The whole process is detailed in Appendix A. Run time critical components of the group lasso solver are implemented in C for fast evaluation. The online operation consists of explaining words with Wikipedia concepts and logging user activity and feedback.

The component providing computational linguistics resources from Wikipedia is a Python library. It is implemented on top of the Java based industry standard Lucene⁷ information retrieval framework. The transition between the languages is done by JCC⁸. This component indexes the 40+GB XML dump of Wikipedia, and provides an easy-to-use API. Among its features are gathering concepts based on a text fragment in term vector representation and fetching contexts for words with a given meaning.

4 Results

In this section, we present the results. Our first result is the methodology to achieve our goals. Our second result is the documentation of the initial activities on the portal.

⁷<http://lucene.apache.org/>

⁸<http://lucene.apache.org/pylucene/jcc/index.html>

The heterogeneous community of robot assisted surgery is in need of reliable information. Our aim was to build a portal system to catalyze the growth of the collaboratively edited knowledge base.

According to [2], the community and the portal system can be considered as a unit called “the global brain”. Our task is its “programming”: determining the flow of information between the community and the machine components.

We discuss the main attributes of the components (including the community), and the features of our system in terms of interactions between the components.

4.1 Components

The system consists of three main components: the knowledge representation, the community, and the artificial intelligence. All three of them are essential. The community needs the knowledge representation for their daily work (in the case of experts) or to access information about, for example, an illness a relative has (in the case of non-experts). The expert community keeps the knowledge base up-to-date with the help of artificial intelligence. Artificial intelligence facilitates understanding by the public and cooperation between different disciplines by explaining the meaning of technical terms. It also automates some processes to extend the knowledge base.

4.1.1 Knowledge Representation

The knowledge representation contains two types of knowledge: a formal ontology and texts in the form of wiki documents. Both are useful to the community and the artificial intelligence.

The ontology provides an unambiguous, formal definition of concepts in robotic surgery. The community can easily see the relations between concepts and navigate their graph along these relations. As we use a Semantic Wiki, the ontology is easy to edit: it is a part of the text. The system also has a full-fledged ontology in OWL for inferencing. At the moment, this ontology is edited separately, and no consistency check between the two are performed, which is a field of ongoing research.

The texts – similarly to Wikipedia – define concepts in natural language. They are a collection of wiki documents – easily manageable by the members of the community. A wiki document is structured text with links to other

wiki documents or third party resources. Besides the community, these documents are also used by the artificial intelligence as training examples for word explanation (Appendix A).

To increase the quality of information in the knowledge base, we introduced two kinds of wiki documents, with different trust levels:

1. *Draft*

The most common document type, as it can be edited by any expert of the community. There is no guarantee that its content is verified, although the system provides a discussion page for quality control.

2. *Article*

A draft can be promoted to an article when it matures. Articles are authored by highly qualified experts agreed upon by the contributors of the draft and approved by an editorial board member (see Section 4.1.2). The starting point of an article is the text in the draft, but it connects only loosely to it – it is about the same topics. The content is verified by a high-quality peer-review process. Articles are fully citeable scientific papers.

4.1.2 Community

The system is based on perpetual contribution by the community and the synergy between their collective intelligence and the knowledge base. The contributors develop the collective knowledge representation by continuously sharing and merging their personal knowledge in their field. They also provide feedback to improve the artificial intelligence. Community is the central component of the system.

Our aim is to form a stable, self-sustained community in robotic surgery, and to provide continuously enhanced and updated information for them. As the community is the central component, we provide additional incentives besides the up-to-date high quality knowledge base.

Experts of robotic surgery are from a wide range of domains: surgery, robot engineering, business, jurisdiction, human and machine cognition, neuroscience, etc. The incentives provided by the system should handle this high level of variety.

The system incorporates appropriate motivation for every member of the community. Experts from the academic sphere can gain reputation by pub-

lishing peer-reviewed, citable articles. They can increase their social capital by taking part in discussions with other members of their field. Firms interested in robotic surgery gain wider recognition by introducing their products (e.g., robots for surgery) to the knowledge base.

Trustworthiness of the knowledge base is maintained by assigning different contribution roles for differently trusted community members. Roles and their level of authority are assigned as follows:

1. *Contributor*

Each registered expert becomes a contributor. Contributor rights can be revoked in the case of abusing them.

- Every *contributor* can start authoring a draft and edit drafts started by others.
- She can propose and vote for a possible author for the article in the topic of any draft she has contributed to.
- She can use every feature of the system facilitating communication and understanding, like a forum or the Word Explanation feature.

2. *Editorial board member*

Similarly to scientific journals, the system has an Editorial Board.

- Members of the editorial board have all rights assigned to *contributors*.
- She decides on the authors of articles.
- She has the right of rejecting any draft appearing in the knowledge base.

Section 4.2 gives details on the procession of contributions.

4.1.3 Artificial Intelligence

Artificial intelligence facilitates understanding by the public and cooperation between different disciplines by explaining the meaning of technical terms. It also automates some processes to extend the knowledge base. Machine learning algorithms are applied on the collected data to support the collaborative work of the community.

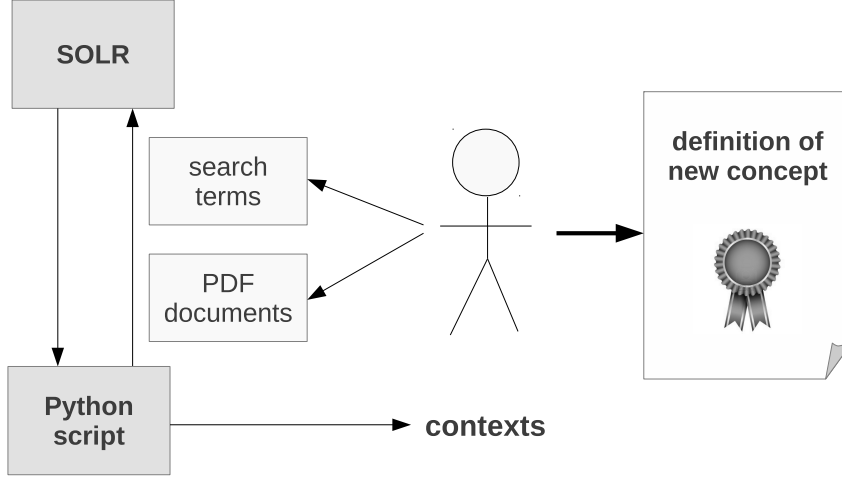


Figure 2: The workflow of extending the ontology. For every new concept to be entered into the ontology, experts process a document collection with our information retrieval system. They provide search terms and expressions, which are used to fetch contexts from the documents.

The communication between sub-domains of robotic surgery is helped by the word explanation feature. Given a word, it provides a few concepts and their short definition from Wikipedia. It can be used by a contributor who encounters a terminological term she is not familiar with, or by a non-expert visitor, seeking reliable information. Details on the operation of the word explanation feature are provided in Appendix A.

An information retrieval system is set up to help ontology expansion. Experts augment the knowledge base by adding new concepts and their short definition. For each new concept, an expert determines a set of related documents and processes it with the information retrieval system. She also provides words and expressions that have the same meaning as the new concept, so the system can search contexts for them. The definition of the new concept is written based on these text fragments. The whole procedure is presented in Figure 2. We are planning to develop more intelligent features supporting work with the ontology as discussed in Section 5.

4.2 Interactions

In this section we consider the interactions between the three main components: the community, the knowledge base, and the artificial intelligence. We describe the processes that help sustain the community, and the mechanisms which support the collaborative workflow of developing of the knowledge base.

Each pair of components has a supportive, bidirectional relation between them (Figure 3). The knowledge representation holds the up-to-date and trustworthy knowledge of the community, which contributes by improving it in the reverse direction. The community improves and supervises the machine learning algorithms in the AI by creating and verifying content, and is helped in this workflow in return. The AI algorithms depend on the knowledge representation to function, and can improve it by detecting inconsistencies and adding more knowledge in an automated manner. Along these interactions the whole system – including the collected knowledge, the community and the learning methods – can improve with time.

Since the community is the central component of the system, we explain the synergy between the components from the view of community members. We describe how they integrate the provided features into their workflow to achieve their goals.

4.2.1 Interaction between the community and the knowledge representation

Interaction from the *community towards the knowledge representation* involves two main progresses: (1) collecting knowledge collaboratively and (2) creating reliable resources.

There are two kinds of collaborative workflows depending on the knowledge representation: the community can extend the ontology and edit articles.

Community members can explore and extend the ontology with new concepts in their domain. Furthermore, they can correct the present ontology along their personal knowledge. These functionalities are available through a convenient interface facilitating the collaborative development.

Collaborative interaction arises in a more sophisticated way by evolving textual representation of knowledge. Writing articles collaboratively requires high coordination of contributors [5]. The presented system accomplishes

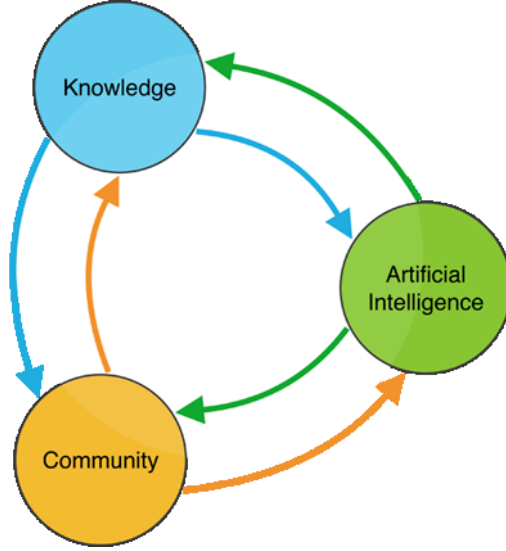


Figure 3: Interactions between the components. Each pair of components has a supportive, bidirectional relation between them. The knowledge representation holds the up-to-date and trustworthy knowledge of the community, which contributes by improving it in the reverse direction. The community improves and supervises the machine learning algorithms in the AI by creating and verifying content, and is helped in this workflow in return. The AI algorithms depend on the knowledge representation to function, and can improve it by detecting inconsistencies and adding more knowledge in an automated manner. Along these interactions the whole system – including the collected knowledge, the community and the learning methods – can improve with time.

both explicit and implicit coordination. In explicit coordination, contributors work out the content through communication. In implicit coordination, a core group of contributors creates the structure of the articles by cooperating, while the majority edits the content along the intention of this group. Explicit coordination realizes in various forms in the system. All wiki drafts in *SurgePedia* have a discussion page like in *Wikipedia*, where the conversation about the topic of the draft can take place. In this way, different parts of the draft page can be planned, or potential mistakes can be discussed and corrected. To reveal lack of topics and discover missing concepts in the knowledge base, there is also a forum and a chat as general communication platforms. From a forum topic or a chat conversation, later a plan or a draft can be crystallized. Implicit coordination is present at editing drafts. Each contributor can take part in the writing process, and the content can be created dynamically.

Reliability and accuracy of the textual knowledge is fostered on three different levels. The first level is a *voting process*. Contributors can vote for promotion of a draft to become an article, when the quality has reached a proper level according to its contributors. The second level is the *editorial board*, who can make decisions about accepting or rejecting articles, and has the right to invite authors to write the *article* about a topic. The last level of quality assurance is a *peer-review* method, which provides journal quality for the published articles by the refereeing of other experts in the field.

Every contributor gets a *contribution score* in proportion to her added value to the page. Contributors, who have a contribution score for the draft (have edited the page), can start a promotion poll and cast a vote. The promotion process consists of three states (states and transitions are illustrated in Figure 4):

1. *Vote on promoting to article*

Each contributor, who has edited a draft, gets a certain amount of contribution score in respect to their contribution to the page content. They are called *draft contributors* of the specific draft. Only draft contributors are able to initiate a poll. After someone started the process, all draft contributors can vote, if they consider it appropriate. Votes are weighted with the contribution score, so the draft will be promoted only if the rate of weighted votes and cumulative contribution score has reached a specific threshold. If it hasn't, draft won't be promoted after a deadline (which can be assigned by the editorial board).

2. *Vote on author of article*

After a draft has passed a promotion poll, contributors have to recommend authors for the to be written article. They can vote on the potential author – chosen from the authors of the references of the draft – by the weighted vote system, described above. The promoted author will be the one who has the most weighted votes from draft contributors. If there is no vote after a deadline, the voting process ends without promotion.

3. *Editorial board decision*

After the vote on the author is completed, the editorial board will be notified of the promotion – there is a list of promoted drafts on a Special Page in the wiki for each member of the editorial board. In this state, the editorial board can accept the draft as an article or reject it. In the latter case, they can remove the promoted draft from the list and the process can restart. In the case that the draft is accepted, the editorial board can propose authors (paying respect to the recommendation of draft contributors or not) to write a journal article into the wiki. After the acceptance, the promotion will be removed from the list, and the draft, related to the article, can evolve further.

4.2.2 Interaction between the community and the AI

Seamless communication between experts are essential for the high coordination of the collaborative work. In our community, there are members from different domains as described in the former section. To channel between different fields, the presented methodology and system provides assistant technologies. For interdisciplinary communication it is not enough to have reliable source information, but we should take into account the difficulties due to the unknown terminological terms across domains. In addition the developed system is aimed to be used also by the wider society of patients, relatives, etc. Non-experts need trustworthy but also understandable information especially in surgical and juridical topics.

To build a bridge between surgeons and robot engineers, neuro-scientists and jurisconsults or surgeons and patients, the presented system provides the formerly introduced *word explanation* method. This process supplies the interaction between the *community and the AI* components. In all textual

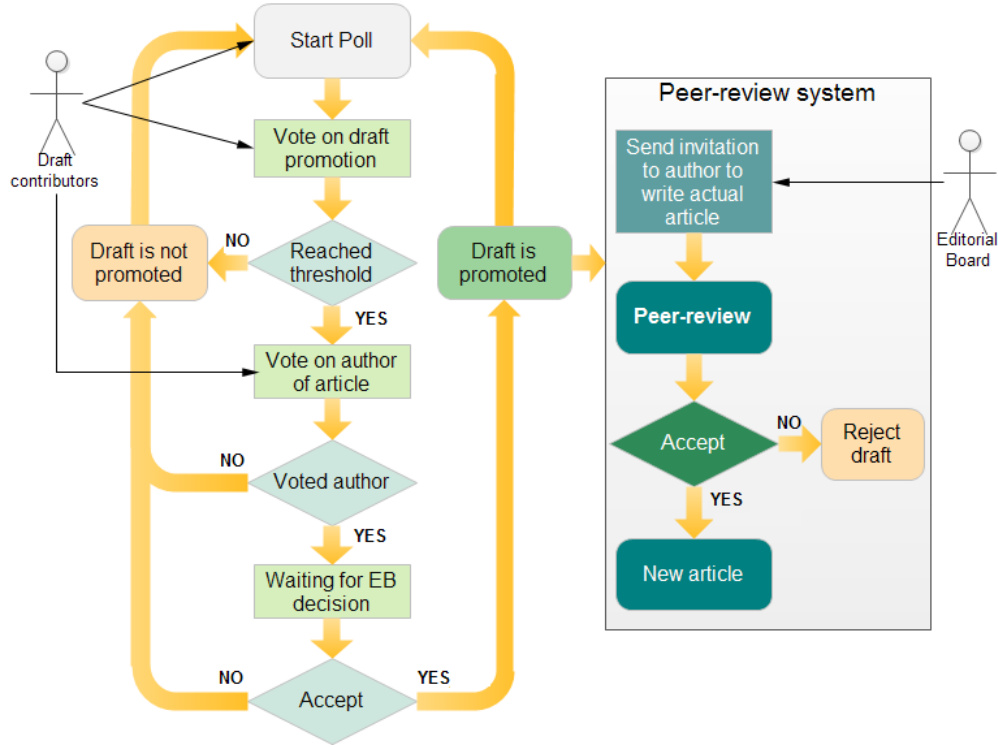


Figure 4: The process of promoting a draft to an article. Each draft contributor can start a poll. The process consists of two voting states, the editorial board (EB) decision state and a peer-review state. If the proportion of weighted votes – either at the state of the article promotion or of the author vote – hasn’t reached the threshold, draft won’t be promoted to article. Otherwise, the process gets into the EB decision state, where members of the editorial board can accept or reject the promoted draft. After the decision, a new draft can be promoted by starting a new poll.

content that appears in the system, words and phrases can be queried through a convenient graphical interface. The word explanation component gives some concepts and their short definition according to *Wikipedia*, fostering the interpretation of the selected terminus. Supporting the interaction in the reverse direction (from community to AI), is a voting feature. Experts of a domain, who have an idea about the selected word or phrase, can cast a positive or a negative vote to each explanation proposed by the AI. This feedback is stored in the system for further analysis and to provide rewards

for reinforcement learning methods, which are the subject of further research.

4.2.3 Interaction between the AI and the knowledge representation

The third aspect of the interactions of system components is the *relation between AI and the knowledge representation*. The connection between the two components can be regarded from two different aspects. First, they are connected through the community, since it is the core of the system, and all the other components are highly dependent on the needs and sustenance of it. Second, NLP methods depend on the knowledge representation to function, referring to the corpora which feed the algorithms. In the reverse direction, results of the algorithms can be integrated into the knowledge representation also in an automated way (for example by the feedback of community).

We focus on the relationship through the community, because in the initial life-stage of the system, building the community and studying the work of the supporting mechanisms are the primary issues. The word explanation feature helps its users elaborate the initial content of the drafts to get into the developing process faster and more efficient. So the concepts of the ontology can be extended, discussions can be started and new drafts can be born to become further articles.

4.3 Initial activities on the EuroSurge Portal

At this stage, the portal is being used by a small community: members of the EuroSurge consortium and invited experts in the field of robotic surgery. Our current aim is to improve the usability and the design of the system before it becomes available to the whole European community of robotic surgery.

Since the portal went online, the experience of this small community was very positive. The knowledge base is growing at a steady pace. More than 20 new drafts were written about concepts that are important in the field of robotic surgery, but not present in Wikipedia. An ontology was developed that provides a formal, unambiguous definition of 100 concepts, including these 20. The ontology is already linked to Surgepedia and Wikipedia to provide easily accessible and clear explanations for these concepts. We expect that when the site opens to a wider public, the knowledge base will grow at an ever-increasing rate.

The drafts of three scholarly articles was written totaling more than 20'000 words. These drafts are continuously being improved upon by the community and – after a peer-review process – can be transformed to articles that contain verified knowledge.

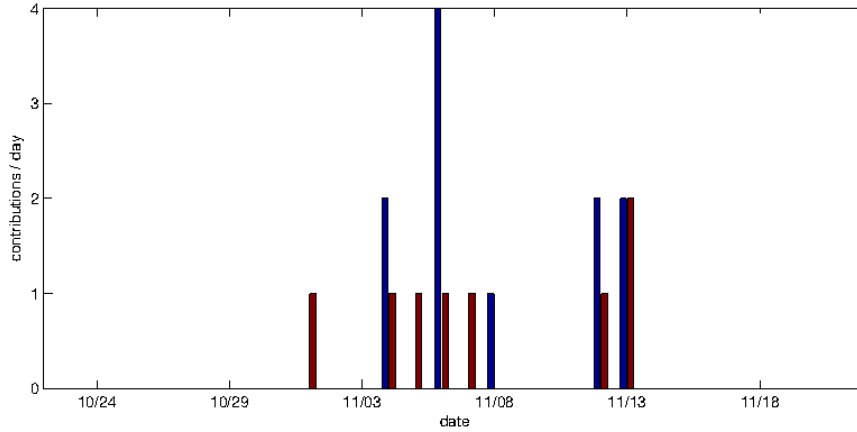


Figure 5: Timeline of a discussion on the portal

A lively discussion took place about the definition of symbols and meaning between two contributors (Figure 5). Based on this discussion, a draft is being written that can be promoted to an article.

These experiences confirm that the system can achieve its goals.

5 Discussion and planned work

We plan to improve upon the system in four key directions: self-sustainability, quality control, creating content, and access to the gathered knowledge.

We will improve the self-sustainability and quality control of the system by introducing a selection mechanism to select and change members of the editorial board based on empirical criteria. In the case of educational games, one such criterion could be the success in predicting the development of the most important users of the portal – children.

Artificial intelligence will also be used for quality control. Currently we can explain the meaning of words by articles from Wikipedia and Surgepedia. The same algorithm can determine whether a word is linked to the correct

article or concept in the ontology. If the algorithm assigns a very low probability to the concept the word is linked to, then we mark the link as probably incorrect.

A central point of interaction between users and the system is content creation. There are already some phases in this workflow that have been automated (Section 4.1). In the future, we will automate most of these phases. We expect that the knowledge representation will grow at a faster rate due to the easier content creation helped by AI algorithms.

On the other hand, content will be collected automatically, by the AI. A focused web-crawler will collect millions of documents related to robotic surgery from the Web, and the AI will filter out the relevant information. This information then will be used to improve the performance of the AI, and to recommend content for experts. The suggested content will help extend the knowledge base.

We plan to provide a wider access to the gathered knowledge by improving the algorithms for word explanation, and by providing additional tools to facilitate better understanding of the difficult concepts encountered in robotic surgery. These tools could incorporate information in the ontology to help explain the unknown concepts in terms of their relations to other concepts.

6 Conclusion

We developed a methodology to solve the problem of building an accurate and evolving knowledge repository from the collective intelligence of experts. We have developed (a) a system that brings together an expert community, who edit the data collaboratively; (b) a methodology and a portal system for the management of a community's collective intelligence with the support of social computing and Natural Language Processing (NLP) methods. Reliability was achieved by distilling the collective intelligence into scholarly papers through voting by the experts, selection by an Editorial Board, and a high quality peer-review process. Our community aims self-strengthening by recruiting experts from Wikipedia and spreading domain information to experts of different fields: terminological terms are explained by machine learning algorithms. We showed how the system works in practice by analyzing the initial activities of the community upon launching.

References

- [1] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.
- [2] Abraham Bernstein, Mark Klein, and Thomas W. Malone. Programming the global brain. *Commun. ACM*, 55(5):41–43, May 2012.
- [3] Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.
- [4] Johannes Hoffart, Torsten Zesch, and Iryna Gurevych. An architecture to support intelligent user interfaces for wikis by means of natural language processing. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, WikiSym '09, pages 12:1–12:10, New York, NY, USA, 2009. ACM.
- [5] Aniket Kittur and Robert E. Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, CSCW '08, pages 37–46, New York, NY, USA, 2008. ACM.
- [6] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM*, pages 233–242, 2007.
- [7] B. Pintér, Gy. Vörös, Z. Szabó, and A. Lőrincz. Explaining unintelligible words by means of their context. In *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, Barcelona, Spain, 15–18 February 2013. (accepted).
- [8] Martin F. Porter. *An algorithm for suffix stripping*, pages 313–316. Morgan Kaufmann Publishers Inc., 1997.
- [9] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.
- [10] Peter D. Turney and Patrick Pantel. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, January 2010.

- [11] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

A Word Explanation explained

Word explanation is an important and mathematically involved component in our system. It is important, because without it, the content would be impenetrable to the public, and – because of the interdisciplinary nature of robotic surgery – to some of the experts as well. It is mathematically involved as it uses advanced algorithms from the field of structured sparse coding. These are necessary because of the inherent complexity of the problem. In the following, we describe the component along the lines of [7].

Our goal is to explain the meaning of a word with a couple of Wikipedia articles. The problem of Wikification [6] - enriching text documents with links to Wikipedia articles - is very similar to our task. Wikification consists of two phases: *link detection* and *link disambiguation*. The detection phase identifies the terms and phrases from which links should be made. The disambiguation phase identifies the appropriate article for each detected term to link to. For example, the term *bank* could link to an article about financial institutions, or river banks.

We consider link disambiguation as our starting point to explain words. Most methods disambiguate words by applying the *distributional hypothesis*: words that occur in the same contexts tend to have similar meanings [3].

There are two main differences between link disambiguation and our problem:

1. We assign *multiple articles* to each word explained.
2. We select the articles from a much larger pool - thousands of candidate articles. In link disambiguation, the problem is simpler: only the candidate articles for a *single word* (i.e., the articles the word links to in the training data) are considered. We do not want to introduce such artificial boundaries for explaining words. The synonyms, hyponyms, hypernyms of a word can be very useful to understand its meaning. We explain the meaning of the word by considering all the candidate articles at once and rely on the distributional hypothesis to select the correct ones.

As the disambiguation problem becomes large scale, exceptions to the distributional hypothesis can occur more frequently. Particularly, let us call two contexts *spuriously similar* if they are similar but belong to different articles. The amount of spuriously similar contexts tends to increase inherently with the number of candidate articles. This makes the learning problem considerably hard.

We address this problem by using the *distributional hypothesis* in a novel way. Structured sparse coding [1] is introduced to diminish the effect of spurious similarities of contexts by utilizing the structure of semantic space (Section A.1).

A.1 The Method

We start from a list of candidate articles the word could be explained as. For each *candidate article*, we collect a number of *contexts*. A context of a candidate article consists of the N non-stopword words occurring before and after the anchor of the link that points to the article. There can be at most $2N$ words in a context.

The presented method makes use of a collection of such *contexts* arranged in a word-context matrix $\mathbf{D}$ [10] (Figure 6). In this matrix, each context is a column represented as a bag-of-words vector $\mathbf{v}$ of word frequencies, where v_i is the number of occurrences of the i th word in the context.

	Boot				Foot				...	Booting			
computer	0	0	0	0	0	0	0	0		1	0	2	1
leg	1	0	0	1	0	1	3	0		0	0	0	0
shoe	0	2	0	1	0	0	0	1		0	0	0	0
:													
modern	0	1	0	1	0	0	0	1		0	1	2	1

Figure 6: The word-context matrix $\mathbf{D}$. Each column is a context of a candidate article (e.g., *Boot*, *Foot*). Each element D_{ij} of the matrix holds the number of occurrences of the i th word in the j th context. For example, the word *leg* occurs three times in the 7th context, which is the 3rd context labeled with *Foot*.

To interpret a word, we formulate an inverse problem: we compute a representation vector $\boldsymbol{\alpha}$ by approximating its context $\mathbf{x} \in \mathbb{R}^m$ linearly with the columns of the word-context matrix $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n] \in \mathbb{R}^{m \times n}$, called the dictionary in the terminology of sparse coding. The columns of the dictionary contain contexts, each labeled with the candidate article $l_i \in L$ the context was collected for. Please note that multiple contexts can be, and in many cases are, tagged with the same candidate article: $l_i = l_j$ is possible. There are m words in the vocabulary, and n contexts in the dictionary.

The representation vector $\boldsymbol{\alpha}$ consists of the coefficients of a linear combination

$$\mathbf{x} = \alpha_1 \mathbf{d}_1 + \alpha_2 \mathbf{d}_2 + \dots + \alpha_n \mathbf{d}_n. \quad (1)$$

For each target word, whose context is $\mathbf{x} \in \mathbb{R}^m$, a representation vector $\boldsymbol{\alpha} = [\alpha_1; \alpha_2; \dots; \alpha_n] \in \mathbb{R}^n$ is computed.

To diminish the effect of spurious similarities, we introduce a structured sparsity inducing regularization by organizing the contexts in $\mathbf{D}$ into *groups*. Each group contains the contexts annotated with a single candidate article. As only a single candidate is selected for each target word, ideally only a single group should be active in each representation vector $\boldsymbol{\alpha}$. Sparsity on the groups is realized by computing $\boldsymbol{\alpha}$ with a *group Lasso* regularization [11] determined by the labels.

The groups are introduced as a family of sets $\mathcal{G} = \{G_l\}_{l \in L} \subseteq 2^{\{1, \dots, n\}}$. There are as many sets in $\mathcal{G}$ as there are distinct candidate articles in L . For each article $l \in L$, there is exactly one set $G_l \in \mathcal{G}$ that contains the indices of all the columns $\mathbf{d}_i$ tagged with l . $\mathcal{G}$ forms a partition.

The representation vector $\boldsymbol{\alpha}$ of the target word whose context is $\mathbf{x}$ is defined as the minimum of the loss function

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{l \in L} w_l \|\boldsymbol{\alpha}_{G_l}\|_2, \quad (2)$$

where $\boldsymbol{\alpha}_{G_l} \in \mathbb{R}^{|G_l|}$ denotes the vector where only the coordinates present in the set $G_l \subseteq \{1, \dots, n\}$ are retained.

The first term is the approximation error, the second one realizes the structured sparsity inducing regularization. Parameter $\lambda > 0$ controls the tradeoff between the two terms. The parameters $w_l > 0$ denote the weights for each group G_l .

If each group is a singleton (i.e., $\mathcal{G} = \{\{1\}, \{2\}, \dots, \{n\}\}$) the Lasso

problem [9] is recovered:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{i=1}^n w_i |\alpha_i|. \quad (3)$$

Setting $\lambda = 0$ yields the least squares cost function.

For the sake of simplicity, we represent each candidate article with the same number of contexts: there are an equal number of columns in $\mathbf{D}$ for each label $l \in L$ ($|G_1| = |G_2| = \dots = |G_{|L|}|$). The weights w_l of the groups are set to 1.

In this group Lasso formulation, whole *groups* are selected. Each group $G_l \in \mathcal{G}$ contains contexts tagged with the same candidate article $l \in L$, and only a few groups can be selected. A context similar to the context of the target word, $\mathbf{x}$, only by accident has a smaller chance to be selected: as it is in a group labeled with a candidate article that is less related in meaning to the target word than the correct candidate, its group contains mainly contexts less similar to $\mathbf{x}$. Therefore, errors introduced by spurious similarities of contexts can be effectively diminished.

A.2 Compiling the dictionary

The dictionary is compiled by sampling the links in Wikipedia. It consists of contexts labeled with candidate articles $(\mathbf{c}_1, l_1), (\mathbf{c}_2, l_2), \dots$. Each labeled context is obtained by processing a link: the bag-of-words vector generated from the context of the anchor text is annotated with the target of the link.

We use the English Wikipedia database dump from July 2012⁹. Disambiguation pages, and articles that are too small to be relevant (i.e., have less than 200 non-stopwords in their texts, or less than 20 incoming and 20 outgoing links) are discarded. Inflected words are reduced to root forms by the Porter stemming algorithm [8].

A list of concepts important for robotic surgery was compiled by experts, and their linked occurrences were collected along with their N -wide contexts. Each link is processed to obtain a labeled context $(\mathbf{c}_i, l_i)$.

To ensure that there are an equal number of contexts tagged with each article $l \in L$, d randomly selected contexts are collected for each label. Labels with less than d contexts are discarded. We do not perform feature selection,

⁹Downloaded from <http://dumps.wikimedia.org/enwiki/>.

but we remove the words that appear less than five times across all contexts, in order to discard very rare words.

B System internals

In this section, we show the inner working of our portal system by presenting use-case diagrams of its features. Figure captions provide details about the software components constituting the system both on the frontend and backend layers. Use-cases are provided for all user roles presented in Section 4.1.2.

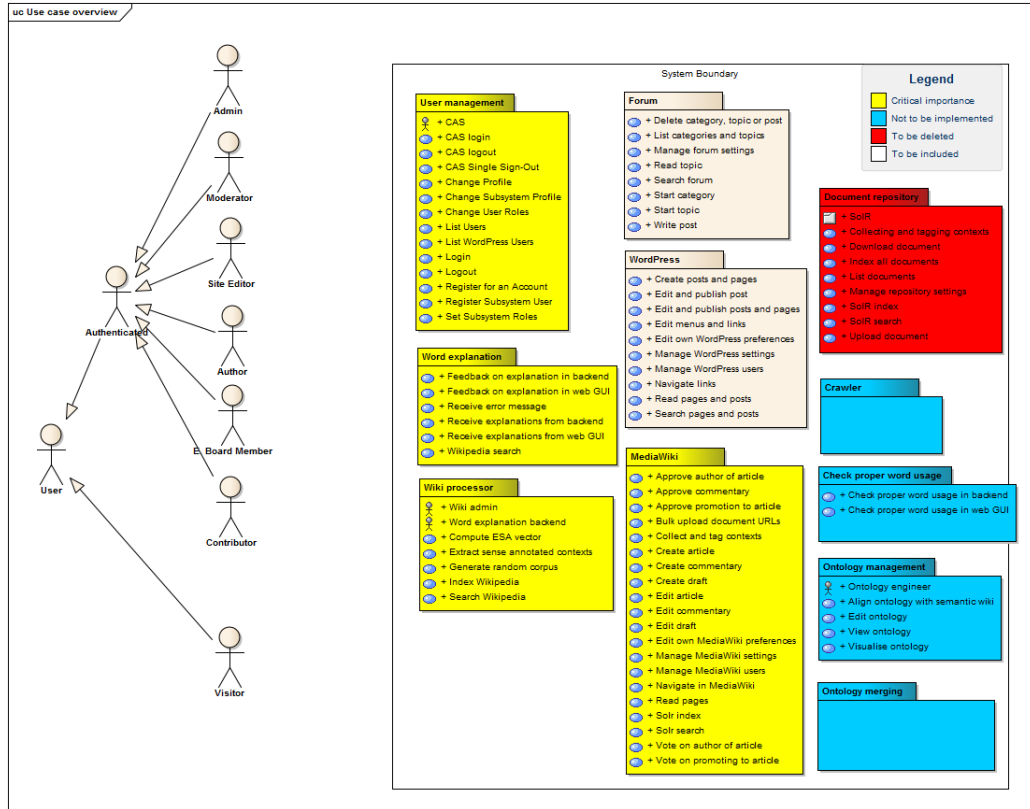


Figure 7: An overview of the portal system. We present the user role hierarchy on the left, and enumerate the use-cases of the portal on the right. Different priorities of features are indicated with different colors.

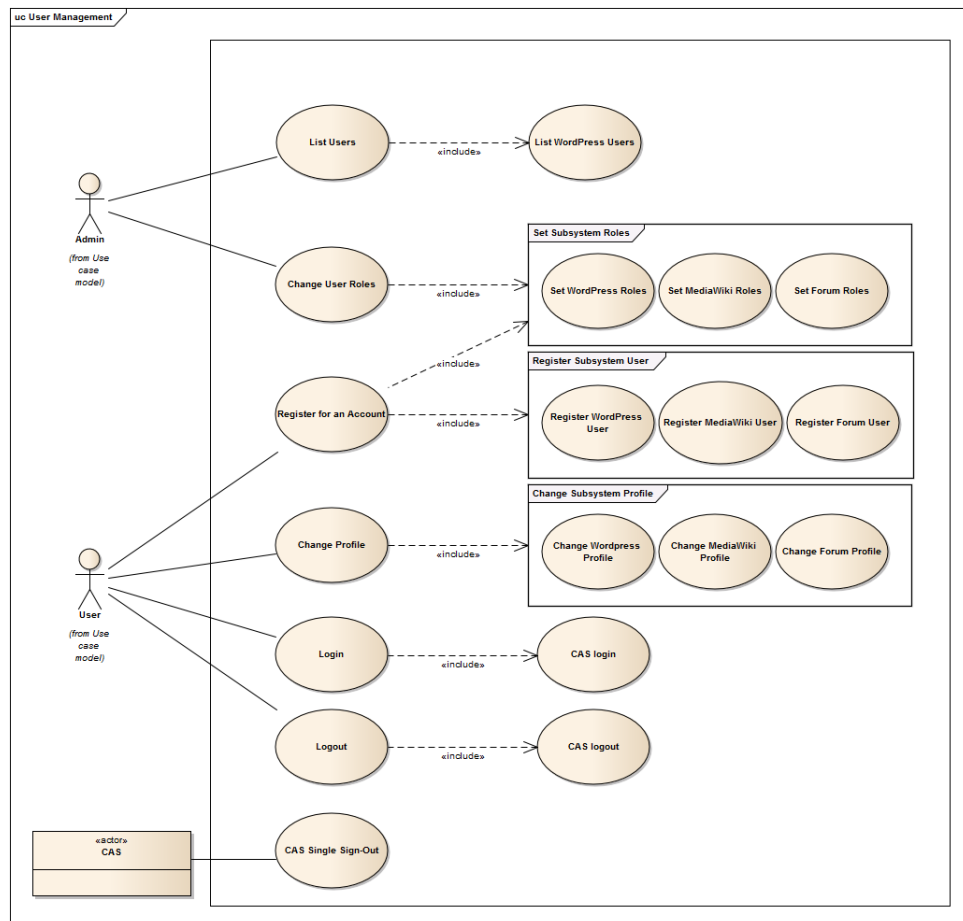


Figure 8: The management of user roles in the early stage of the portal.

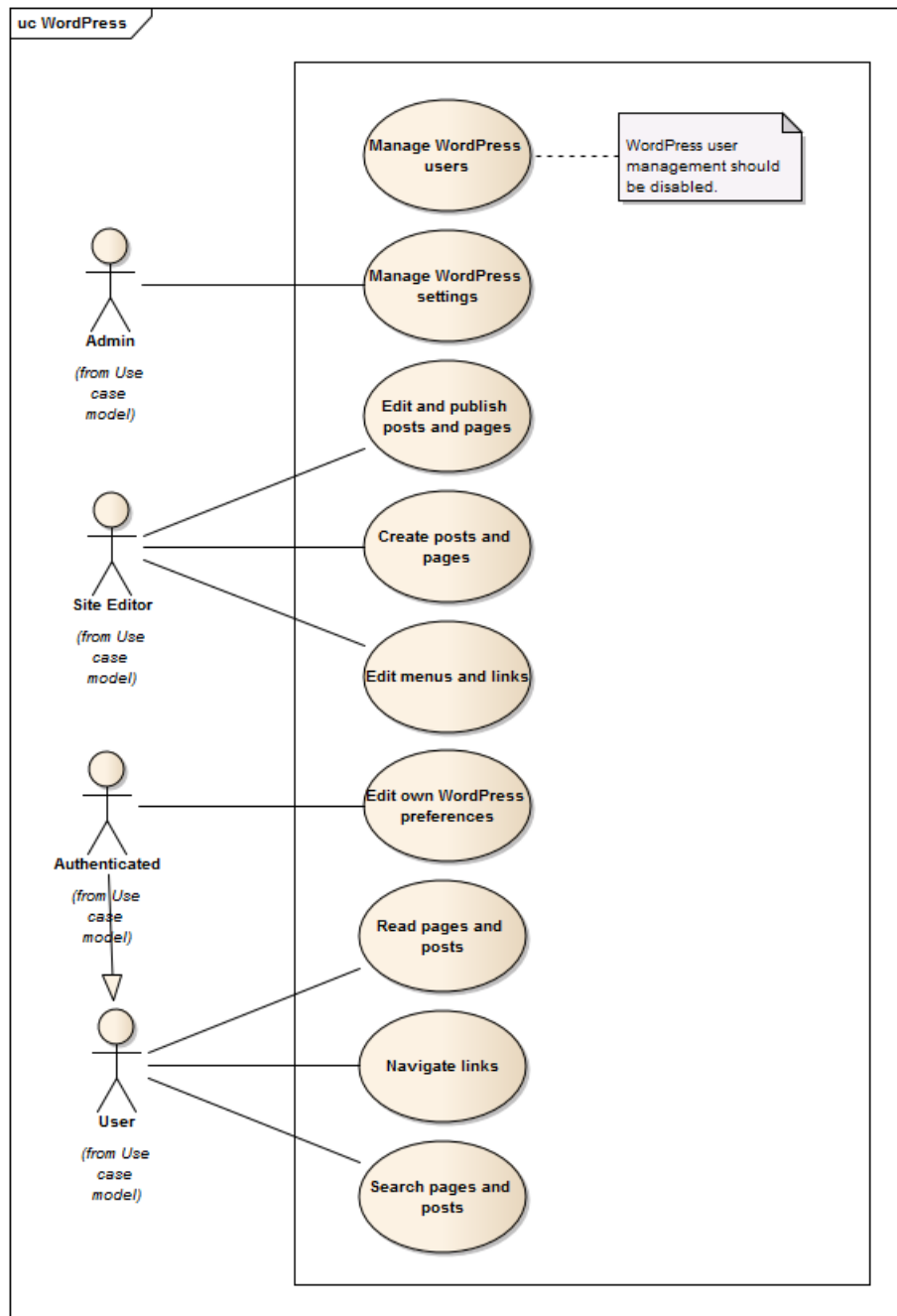


Figure 9: Core use-cases provided by the WordPress component. This popular, open web content management system serves as a news source. It can be extended to provide communication channels like forums or document stores.

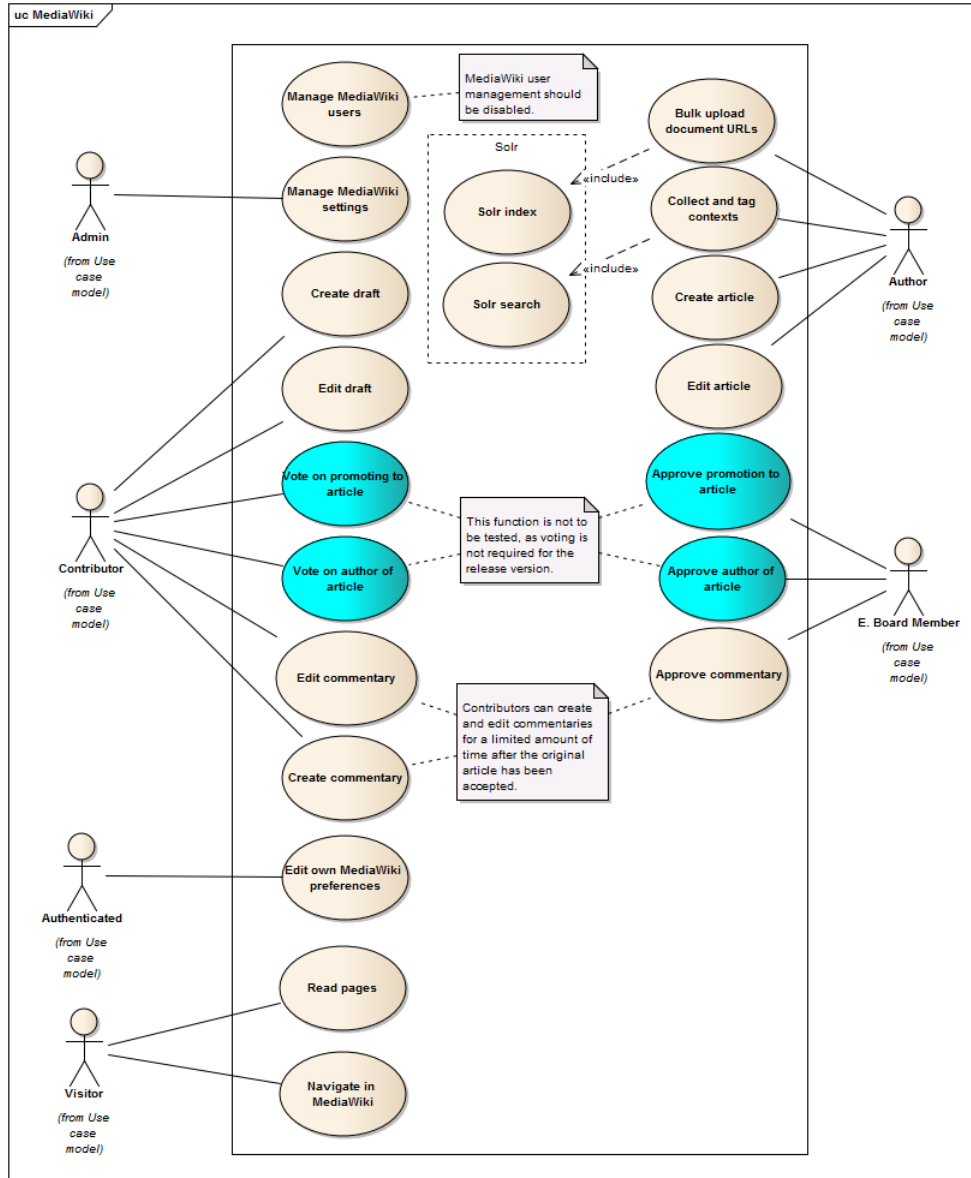


Figure 10: Use-cases provided by the MediaWiki component. MediaWiki provides the Wikipedia-like collaborative content management features. It is extended to support our more sophisticated quality control instruments.

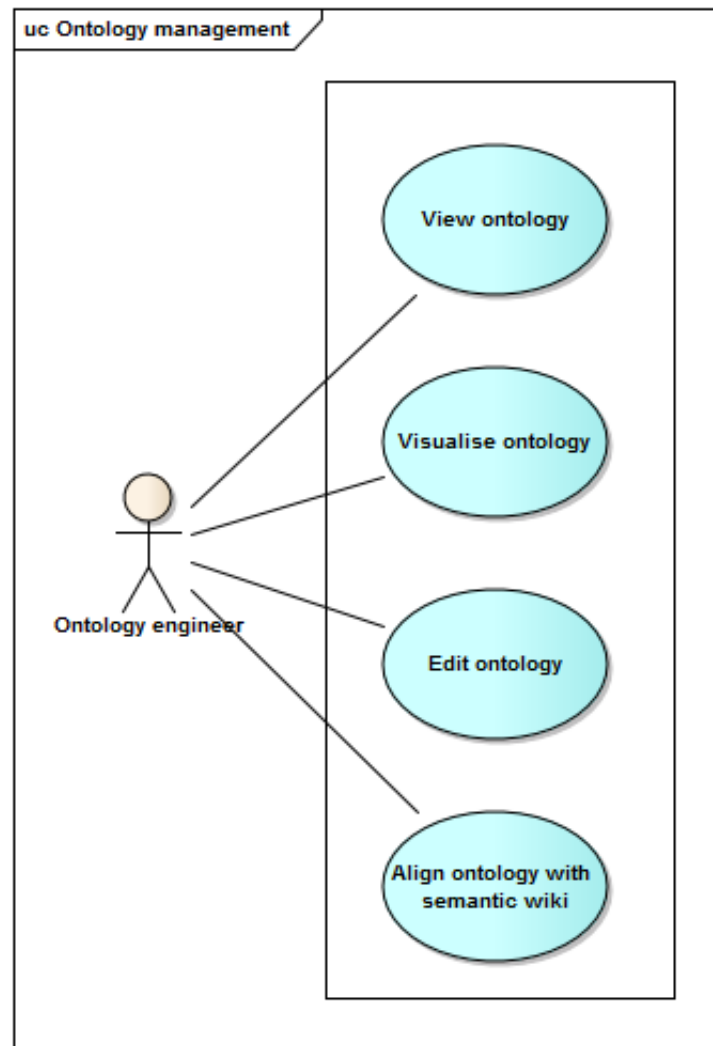


Figure 11: Use-cases related to ontology management. These features are currently handled by WebProtégé. Ontology extension is also supported by an information retrieval system.

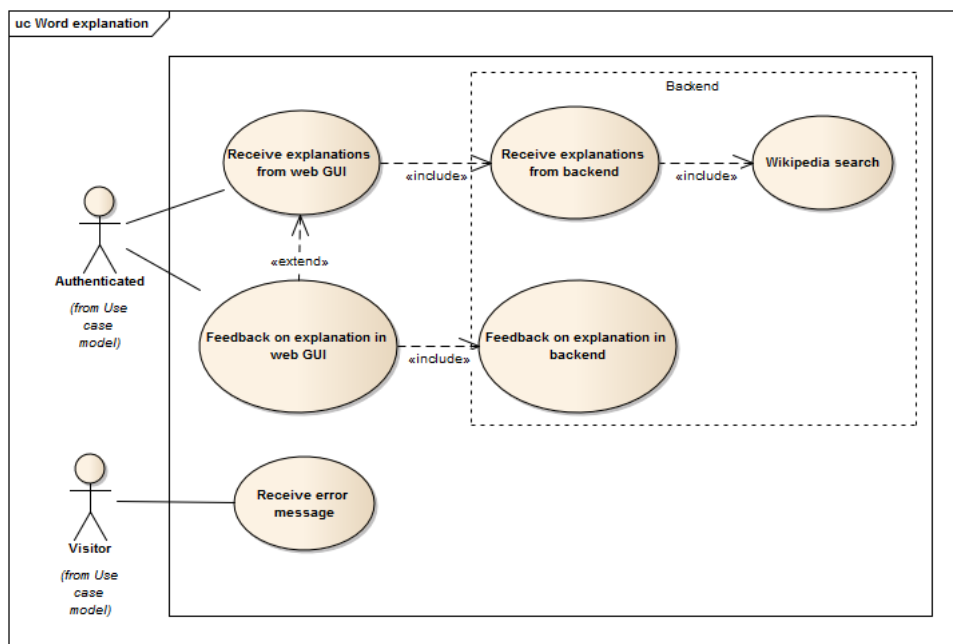


Figure 12: Use-cases of the AI based word explanation feature. The authenticated user gets help in interpreting unknown words, and has the ability to provide feedback on the results.

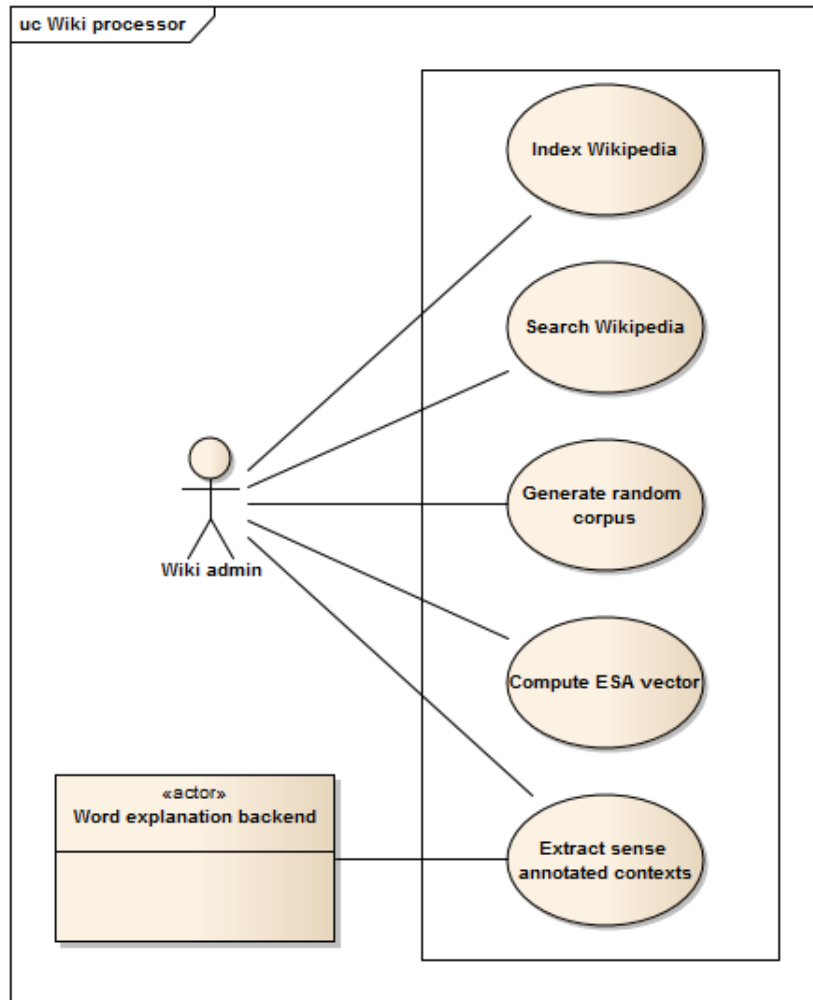


Figure 13: Use-cases of the “wiki processor”. It feeds the word explanation feature with special Wikipedia concepts and related contexts offline.

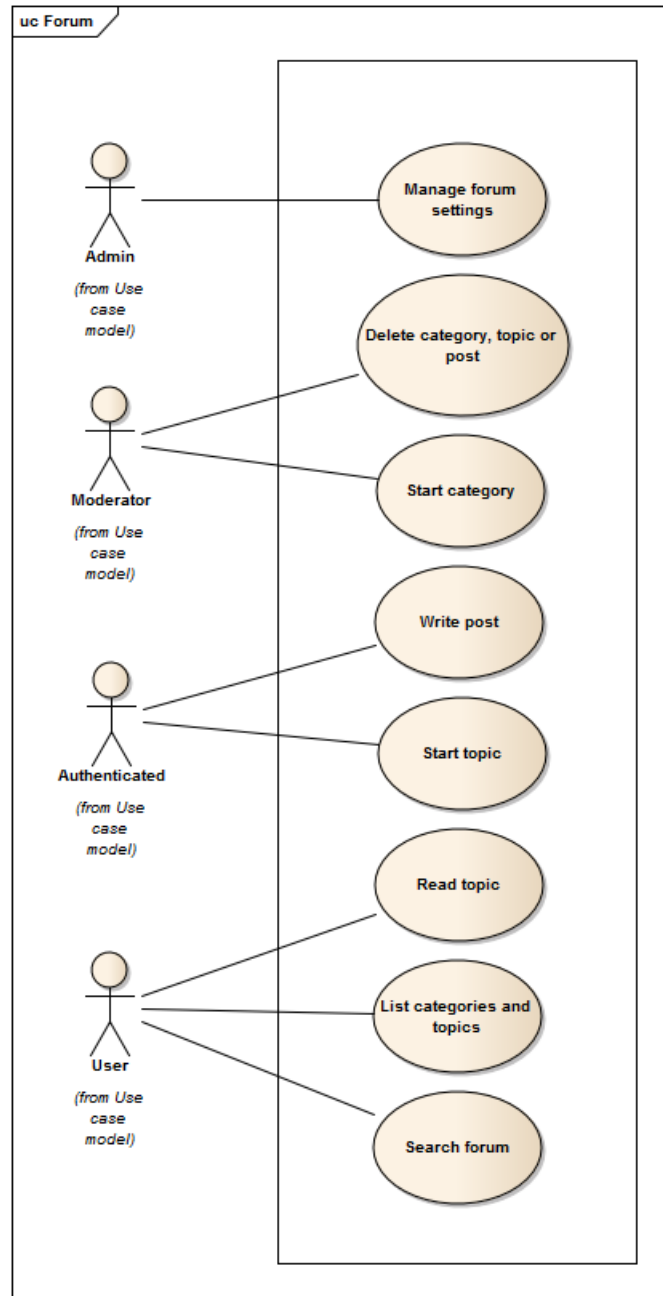


Figure 14: Uses-cases related to forums in the WordPress component.