Available online at www.sciencedirect.com
ScienceDirect
Decision Support Systems 45 (2008) 567–584
Decision Support Systems
www.elsevier.com/locate/dss

Enhancing portability with multilingual ontology-based knowledge management

Aviv Segev *,[1], Avigdor Gal

Technion — Israel Institute of Technology, Haifa 32000, Israel

Available online 6 August 2007

Abstract

Information systems in multilingual environments, such as the EU, suffer from low portability and high deployment costs. In this paper we propose an ontology-based model for multilingual knowledge management in information systems. Our unique feature is a lightweight mechanism, dubbed *context*, that is associated with ontological concepts and specified in multiple languages. We use contexts to assist in resolving cross-language and local variation ambiguities. Equipped with such a model, we next provide a four-step procedure for overcoming the language barrier in deploying a new information system. We also show that our proposed solution can overcome differences that stem from local variations that may accompany multilingual information systems deployment. The proposed mechanism was tested in an actual multilingual eGovernment environment and by using real-world news syndication traces. Our empirical results serve as a proof-of-concept of the viability of the proposed model. Also, our experiments show that news items in different languages can be identified by a single ontology concept using contexts. We also evaluated the local interpretations of concepts of a language in different geographical locations.

Keywords: Knowledge management; Knowledge sharing; Ontology; Context; Multilinguality; eGovernment

1. Introduction

Experiences in developing information systems have shown it to be a long and expensive process. Therefore, once a generic information system has been developed, it is the aim of the developer to make it as portable as possible and the aim of users to deploy it with minimum effort. In some cases, such deployment requires the change of language, which affects the user interface as well as the internal decision making processes. In this work we focus on applications in which a language transfer serves as a main obstacle in adapting an information system to user needs. As a case in point, consider eGovernment applications in the European Union. The EU puts effort into homogenizing its governance procedures to allow easy interoperability. Yet it does so without committing to a single language. On the contrary, the EU values the preservation of local culture (including language). In such applications, the development of an information system that is monolingual will result in low portability and high deployment costs and therefore multilingual information systems seem to be more appropriate.

Recent advances in information system development suggest the use of ontologies as a main knowledge management tool. Ontologies model the domain of discourse and may be used for routing data, controlling the workflow of activities, assisting in semantic annotation of both data and queries, *etc*. In this paper we take advantage of these

* Corresponding author.
E-mail addresses: asegev@tx.technion.ac.il (A. Segev), avigal@ie.technion.ac.il (A. Gal).
[1] Present address: National Chengchi University, Taipei 11605, Taiwan.

0167-9236/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2007.07.011

recent advances and propose an ontology-based model for multilingual knowledge management in information systems. Our mechanism is based on a single ontology, whose concepts can have multiple representations (*i.e.*, concept names) in various languages. While such solutions already exist (*e.g.*, in Protégé), we argue that they are insufficient. On the one hand, a single global ontology is preferred over local ontologies when it comes to interoperability. On the other hand, mere translation of ontological concepts from one language to another is insufficient to fully represent differences that may arise from the change of language. Such differences may result in concept ambiguity and generally in under-specification of semantic meaning [9].

To compensate for ontology under-specification we propose to support multilingual ontologies with a lightweight mechanism, dubbed *context*. Contexts serve in the literature to represent local views of a domain, as opposed to the global view of an ontology [10]. While the specific representation of contexts vary, one may envision a context, as an example, to be represented by a set of words, possibly associated with weights, reflecting some notion of importance. Contexts, in our proposed solution, are associated with ontological concepts and specified in multiple languages. Therefore, they aim at conveying the local interpretation of ontological concepts, thus assisting in the resolution of cross-language and local interpretation ambiguities.

Equipped with such a model, we next provide a four-step procedure of overcoming the language barrier in deploying information systems. We also show that our proposed solution can overcome differences that stem from local interpretations that may accompany cross-language information systems deployment. The proposed mechanism was tested in an actual multilingual environment in the framework of the QUALEG (Quality of Service and Legitimacy in eGovernment) project.[2] The QUALEG project is an innovative project sponsored by the European Union to promote the relationship between local governments and citizens. This multilingual information system aims at allowing local governments to maintain a direct connection with citizens through the ongoing adjustment of their policies according to the assessment of citizen needs. This implies that local governments should be able to measure the performance of the services they offer, assess citizen satisfaction, and re-formulate policy orientations on such elements with the participation of citizens, all in a multilingual setting.

To complement our experiences with QUALEG, we provide a set of experiments performed in a controlled environment using news syndication data. Our empirical analysis shows that news items in different languages can be identified by a single ontology concept using contexts. We also evaluate the local interpretations of concepts of a language in different geographical locations and show that the results drop when using a different local interpretation training set and drop considerably when using a mixed local interpretation training set from an identical language.

To summarize, our main contributions are as follows:

- We propose a knowledge management model, based on the relationships between ontologies and contexts, which lends itself well to the support of effective portability and deployment of multilingual information systems.
- The high degree of flexibility the proposed model provides is translated into procedures for the deployment and querying of a multilingual information system.
- We demonstrate the feasibility of our model using an implementation and deployment in the context of a European eGovernment project.
- We provide a thorough empirical analysis, revealing that joint classification by uniting all languages in a single concept has minimal impact on the results.

The rest of the paper is organized as follows. Section 2 provides a review of related work as a starting point for the ontology-based multilingual knowledge management model, proposed in Section 3. Section 4 provides the methodology for managing ontologies when deploying a new multilingual information system and when querying it. Section 5 describes our experiences with the QUALEG project and experiments with RSS news data. We end with concluding remarks in Section 6.

2. Related work

This section reviews previous research work relevant to this paper. We start by discussing current support for multilingual ontologies (Section 2.1). We then discuss the machine translation approach for multilinguality (Section 2.2), and current techniques of multilingual information retrieval (Section 2.3).

2.1. Ontologies and multilinguality

Ontologies are currently considered the de-facto standard for representing semantic information. Their design, however, is a difficult task, requiring the collaboration of ontology engineers and organization experts. Therefore, ontologies are manually crafted and tuned, which results in a static domain model, infrequently modified.

[2] http://ww.qualeg.eupm.net.

Nevertheless, once designed their universal nature makes them an excellent mechanism for application interoperability. Without universal ontologies, interoperability becomes an uncertain process [8], which may not be acceptable for some applications.

The static nature of ontologies conflicts with the dynamic nature of the world. Businesses nowadays often change and need to adapt the semantic representation of their occupations to their changing business environment. Governments, which change less often, still need to adapt their own regulations to a global community, while maintaining some divergence from standard governance, reflecting local interpretations and lingual differences. The research literature has proposed a hybrid approach [14], in which ontologies are recognized as static entities yet an organization can change its business semantic representation dynamically. To do so, an ontology is defined to have two parts: a static part (which is the global ontology) and a dynamic part, which evolves either by exporting ontologies or by discovery. With such a model at hand, organizations can still interoperate using the universal part of the ontology, and continuously change their business models using the local component of the ontology. Every now and then, an industry may recognize certain practices as universal and add them to the global ontology, a standard accepted by consensus. The model we propose in this work replaces local ontologies with contexts, which are easier to extract and manage.

To support multilingual applications, ontologies can separate their internal concept descriptors (typically semantic-less textual description) from the name of a concept. In doing so, one can associate more than a single name to the same ontology concept. In particular, if each name is given in a different language, one can construct a clear and cohesive system, without redundancy. Such a support has already become a standard and is used by several ontology management systems (*e.g.*, Protégé). We extend this idea into multilingual contexts.

2.2. Translation

The first solution that comes to mind when working with multiple languages is to use translation. However, translation entails language-specific difficulties, such as the importance of the connection between grammar and meaning, the role of word endings and word position, and the length and complexity of words, which are comprised of other words. Translation also entails difficulties that arise from the translation effort itself: some words do not have exact parallels in other languages, nuances are hard to convey, and a word may have different meanings in different contexts.

The use of automatic tools for language translation has been suggested as a solution for multilingual applications [33]. However, this solution is not viable, since automatic machine translation (MT) today suffers from several critical limitations [13]. First, these tools have yet to achieve a level of proficiency comparable to human translation. Although there are no universally accepted evaluation methods, different methods of evaluation of MT in specified operational contexts still indicate that MT does not attain a sufficiently good level, in terms of measures such as intelligibility, accuracy, fidelity, and appropriateness of style. While human translation can identify errors and deficiencies that can be corrected or improved, MT has yet to acquire this ability. A person who makes a mistake once can learn for the future, but MT still cannot. Currently, any prospect of a fully automatic general-purpose system capable of good quality translation without human intervention is beyond the scope of MT.

Therefore, this paper presents a solution that bypasses machine translation in multilingual environments by using a single ontology system to which predetermined manually translated ontology concepts are automatically mapped. We compensate for under-specification of the ontology by using contexts, local view points that can be automatically generated in a language-independent fashion.

2.3. Multilingual information retrieval

Research into multilingual information retrieval has been going on for more than 50 years. In recent years, the increasing impact of the Internet has generated even more interest in the topic [12].

One issue with multilingual information retrieval refers to the performance of information retrieval in different languages. The performance of AlltheWeb, Altavista, Google, and three Arabic engines was examined by Moukdad [21] to see how they handle Arabic linguistic characteristics. He found that it is necessary to make users aware of the limitations of general search engines in retrieving Arabic documents. Along the same lines, Bar-Ilan and Gutman [3] examined the performance of AlltheWeb, Altavista, and Google to handle four different languages: French, Hebrew, Hungarian, and Russian. They reported that "non-English languages have a much larger chance of being lost in Cyberspace". Multilingual information retrieval from the Internet in the Turkish language entails certain problems, particularly due to the existence of special characters in the language [1].

Another line of research involves cross-retrieval among different languages. One approach to cross-lingual text retrieval (CLTR) using multilingual text mining finds the multilingual concept–term relationships

from linguistically diverse textual data relevant to a domain. These, in turn, are used to find the conceptual content of the multilingual text. When language-independent concepts hidden beneath both document and query are revealed, concept-based matching is made possible [6].

Examples of other multilingual systems include MEMPHIS, an agent-based system for enabling acquisition of multilingual content [24], and MUMIS multimedia indexing through multi-source and multi-language information extraction [28].

The model presented in this paper uses ontologies and contexts for multilingual knowledge management, including tasks of information retrieval. Our model is language independent and requires minimal training to define a topic automatically. Furthermore, the model allows cross-lingual storage of information and can integrate some of the aforementioned techniques of information retrieval.

3. A model for multilingual knowledge management

This section presents a model for multilingual knowledge management in information systems. The model is based on a semantic representation tool (ontology) enhanced with contexts. We start by formally defining contexts and ontologies and their inter-relationships (Section 3.1). We then present the use of ontologies and contexts in supporting multilingual tasks (Section 3.2). Finally, we discuss the advantages of using the model for multilingual knowledge management (Section 3.3).

3.1. Preliminaries

A common definition of an *ontology* considers it to be "a specification of a conceptualization" [10], where conceptualization is an abstract view of the world represented as a set of objects. The term has been used in different research areas, including philosophy (where it was coined), artificial intelligence, information sciences, knowledge representation, object modeling, and most recently, eCommerce applications. For our purposes, an *ontology* $O=V,E$ is a directed graph, with nodes representing concepts (vocabulary or things [4,5]) associated with certain semantics and relationships [26]. For example, in eGovernment a concept can be Public Service with a relation includes to a concept Activity of Public Administration and a relation responsibility to a concept Local Spatial Management Strategic Plan. Typically, ontologies are represented using Description Logic [7], where subsumption typifies the semantic relationship between terms; or Frame Logic [15], where a deductive inference system provides access to semi-structured data.

We define a descriptor c from domain D as an index term used to identify a record of information [20]. It can consist of a word, phrase, or alphanumerical term. A weight $w \in \Re$ identifies the importance of descriptor c in relation to the record of information. For example, we can have a descriptor Immovables, and a weight of 6. A *descriptor set* $\{\langle c_i,w_i\rangle\}$ is defined by a set of pairs, descriptors and weights.

By collecting descriptor sets together we obtain a *context*. A *context* $C=\{\{\langle c_{ij},w_{ij}\rangle\}_i\}_j$ is a set of finite sets of descriptors. For example, a context C may be a set of words (hence D is a set of all possible character combinations) defining a document *Doc* and the weights can represent the relevance of a descriptor to *Doc*. In classic Information Retrieval, $\langle c_{ij},w_{ij}\rangle$ may represent the fact that the word c_{ij} is repeated w_{ij} times in *Doc*.

3.2. Ontologies, contexts, and multilingual knowledge management

We now move on to describe a model for multilingual knowledge management using ontologies and context. We can consider each descriptor c to be a different point of view of some concept $v \in V$. A descriptor set then defines different perspectives and their relevant weight, which identifies the importance of each perspective. For example, an ontology concept Local Spatial Management Strategic Plan can be represented by descriptors such as: ⟨Immovables, 40⟩, ⟨Building, 25⟩, ⟨Infrastructure, 20⟩, *etc*. We can now assume that each descriptor set represents a different language and then a context is a multilingual representation of a concept.

The proposed model associates an ontology concept with a name and a context. We extend the multiple-name support mechanism, described in Section 2.1 and propose multiple-context support in a similar fashion. A concept is associated with multiple contexts (note that in [32] we have defined a context algebra that is closed under the union operator and therefore multiple contexts are in themselves a context), each in a different language. Fig. 1 provides a schematic illustration of our model for multilingual knowledge management. Four ontology concepts are displayed: Public Service, Citizen, Activity of Public, and Local Spatial. Each one has concept names also in French, German, and Polish. For the Local Spatial concept, a set of contexts represents the local perspective of the concepts in both English and Polish.

3.3. Discussion

One may argue, and rightly so, that an ontology concept needs no interpretation. All one needs to know about a

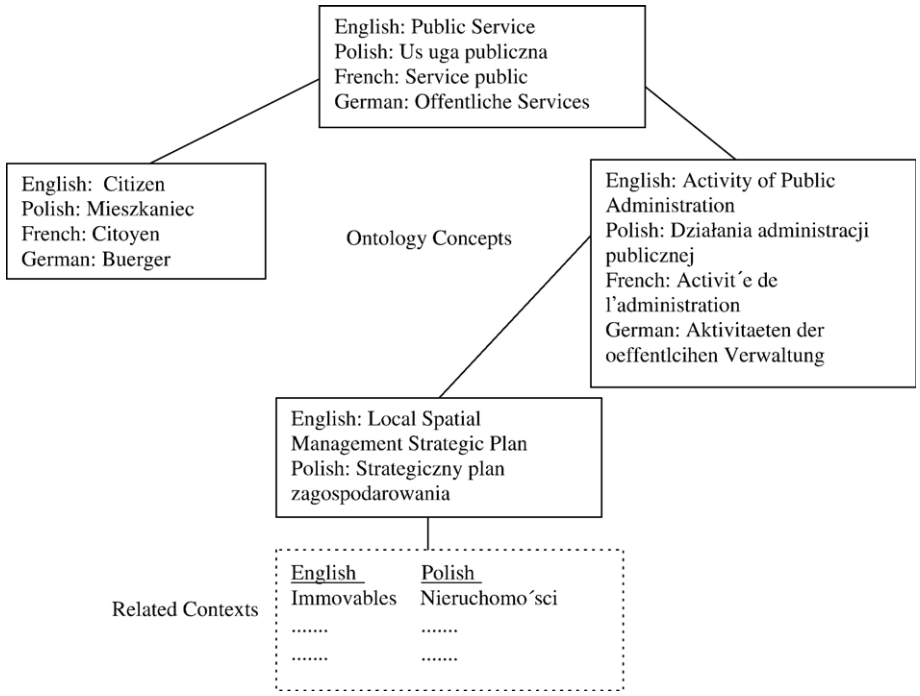

Fig. 1. Multilingual ontology example.

concept can be deduced from the ontology and from reasoning about the semantic relationships between a concept and other concepts. For example, consider the concepts of Public Service and School: Public Service encompasses School since School is an area of responsibility of Public Service, while School is administered as part of a Public Service. This argument holds true as long as the ontology fully specifies the domain of discourse. However, there is a philosophical ongoing debate on whether fully-specified ontologies actually exist. Classical results (*e.g.*, the model-theoretic argument of the philosopher H. Putnam in [25]) indicate that there is nothing to be done to prevent unintended interpretations of clauses in a formal language. Therefore, to be pragmatic, we must assume that a universal ontology is under-specified in that it fails to identify local variations. For example, the concept of Rahmenprogramm in German can be translated into a master program, fringe event, or supporting program but cannot be understood from its relation to the concept of Finanzen (finance). Therefore, a context is necessary to understand a concept whenever an ontology is under-specified and to fill this semantic gap. The contexts associated with Rahmenprogramm include: Pressekonferenz (press conference), Festivalclub, and Festveranstaltung (festival event), indicating that its true meaning here is the festival main program.

Contexts are less expressive than ontologies. Their interpretation of a concept is "flat" in that they describe only a vague notion of semantic relationship and their structure is typically limited to a set of keywords. For example, the synonymous contexts Bulletin, Eintritt (admission), and Startschuss (starting shots) can only assume a clear meaning once they are related to the concept of Rahmenprogramm. Therefore, one may derive a greater benefit from representing the local component as an ontology as well (which was the proposed solution in [14]). However, recall that ontology engineering is a manual and difficult task. This brings us to the main difference between previously proposed solutions and our model. Contexts can be automatically generated and we present, by way of motivation, an algorithm for context extraction in Section 4.2.1. Contexts can also be manually tuned by organization experts, rather than ontology engineers, which enables a more dynamic modification of contexts, better representing the evolving business environment.

We see three main benefits of the proposed model:

Flexibility with respect to a global ontology

Newly developed applications follow some common standard and come, most likely, with a generic global ontology. Such an ontology is the outcome of a well-designed set of concepts and relationships modeled by experts. Once defined, evolving it becomes a difficult task. Here, contexts can serve as the local interpretation of the global ontology, which can be maintained by a local expert without the involvement of the IT personnel. For example,

to utilize an ontology concept of Spatial Management in the public service all the civil servant needs to do is add a context, semantically translating the concept to her local language.

Flexibility with respect to language

Multilinguality results in a need to adapt the ontology to different languages separately. Avoiding such multiple efforts is desirable, both for the initial specification of the ontology and for the ontology evolution. Here, the context can serve as the "translation" mechanism, in which ontological concepts are interpreted in the local language. To illustrate this point, consider the English concept state, representing a medium government level. While Germany uses a translation of the English concept state for this concept, Poland officially uses a related term of a province. The latter seems to represent an entity with less autonomy (similar to the difference between the federal systems in the US and Canada) than the former. When translating the English word state to German, one gets ten concepts that are close in meaning to that of a state, while with Polish, there are fourteen possible meanings. The use of a context and the mechanism suggested below for generating the context of a concept (such as state) compensates for any under-specification that may result from the universality of the ontology.

Flexibility with respect to local interpretations

Even without a language barrier, different entities may have different emphases on this or that task, emphases that represent local interpretations. For example, border cities (*e.g.*, Saarbrücken at the German–French border) may put more emphasis on recognizing language differences than cities in the heart of countries, therefore investing more in multicultural events. Also, capital cities may have more sensitivity to minority culture than cities in the periphery. Context can therefore serve as a compensating element in ontologies, adding topics of interest to the global ontology.

4. From a model to a multilingual information system

Equipped with the model, presented in Section 3, we now illustrate its usage by providing a four-step procedure for deploying a new multilingual information system. The procedure is focused on adapting an existing ontology to the needs of a new information system. It includes the following four steps, *selection*, *collection*, *extraction*, and *adaptation*. Selection involves selecting an existing ontology. In the collection step, sample documents that represent ontology concepts are collected. Contexts are extracted from the sample documents in the extraction step. Finally, extracted contexts are associated with ontology concepts. We now describe each of these steps in more details.

The first step of ontology selection involves a selection of an *existing* ontology, relevant to the information system domain of discourse. Such an ontology can be part of the information system, or can be exported from brokers that specialize in domain-specific ontologies (*e.g.*, ontology.org). If an ontology designer is involved, one may consider designing a new ontology from scratch, revising an existing ontology, or integrating existing ontologies from several domains. One approach to a multilingual ontology can focus on the automatic generation of a hierarchical knowledge map — NewsMap [23], which displayed a technique for extracting relevant phrases from a news collection using a statistical phrase extractor, hierarchical categorization, and knowledge maps visualizer. In [31], a semi-automatic method for supporting organization experts (as opposed to ontology engineers) in the task of evolving ontology was provided, increasing the automation of this step.

In the collection step, an organization expert is requested to provide organizational documents that best describe each concept in the ontology. Sample documents can include organization documents representing the concepts, news articles representing the topic, or correspondence, such as email, between the organization and its clients. This step is the most manual-intensive of all four steps. Existing documentation of document classification in an organization may prove to be useful here. For example, to identify documents for a yearly Theater Festival concept in a local government, documents describing the festival from the previous years can be used.

The extraction step yields a context for each concept in the ontology. The context is computed from the set of documents deemed relevant for a concept. Due to a long research tradition in the area of automated text categorization (see, for example, a survey in [29]), this step can be done automatically. For illustration purposes, we provide in Section 4.2.1 one such automatic context extraction algorithm. We consider our ability to use automatic means for this step a major benefit of our model, as opposed to existing models, *e.g.*, [14], where the local view of the system is given as an ontology, which is hard to design and maintain by organization experts.

The final step is a technical one. It involves adding the new contexts to their relevant concepts. It is worth noting that at this time the organization expert may decide to either keep existing contexts as well, or remove them. Keeping existing contexts can assist in multilingual information systems, where contexts in different languages can co-serve in tasks of knowledge management. However, if these contexts are too greatly oriented towards local interpretations, then keeping existing contexts from other

deployments of the same information systems may harm the system performance.

4.1. Document collection

The document collection step is an essential step in providing the information system with the correct interpretation of ontology concepts. To highlight the benefits of this approach, we continue our discussion along the lines of the three benefits, as proposed in Section 3.3:

Flexibility with respect to a global ontology: Recall that a global ontology serves as a basis for local applications. An organization expert, given an ontology, determines the most appropriate set of documents of each concept, based on the set of concepts in the ontology. Therefore, the domain expert implicitly compensates for ontology under-specification by manipulating the relevance of documents, associating them with the concept that is most relevant in the given ontology.

Flexibility with respect to language: The documents, as provided by the organization expert, are given in the local language. In the discussion in Section 4.2 we emphasize that the extraction algorithm should be language independent, and therefore the generated contexts are given in the local language. At times, if a term in a different language is closely related to the context, it will be added as well. For example, in Saarbrücken at the German–French border the French named Perspective du Theatre festival encompasses German related documents.

Flexibility with respect to local variations: Local variations will be taken into account, using the local organization expert document classification. Therefore, if a certain document falls under one concept in one organization and under another concept in another organization, such classification will affect the context generation process.

This step relies, to a great extent, on the subjective assessment of a local organization expert. Therefore, it may generate undesirable biases in concept interpretation. In the next section we demonstrate how such biases can be countered by using query expansion for context extraction and recognition.

4.2. Context extraction and recognition

We now focus on the third step, the extraction step. A large body of research exists for extracting context from text. A class of algorithms were proposed in the IR community, based on the principle of counting the number of appearances of each word in a text, assuming that the words with the highest number of appearances serve as the context. Variations on this simple mechanism involve methods for identifying the relevance of words to a domain, using methods such as stop-lists and inverse document frequency [29]. In this section we first discuss the desirable properties of context extraction for our research, followed by a description of a context recognition algorithm, to illustrate our approach. Other models, such as [18] and [11], can be adopted for context recognition as well. Evaluating the best extraction algorithm for our needs is beyond the scope of this paper. We settle for an algorithm that shows reasonable results as a proof of concept. Our experiences and experiments with the context extraction algorithm are detailed in Section 5.

A desirata for an algorithm that implements the extraction step should consist of three elements. First, it should be automatic, ensuring a quick deployment. Second, it should be language independent. This way, it can be applied with each new deployment, regardless of the language of choice. Third, it should circumvent biases that may have been introduced in the collection step, given that the collection step is manual labor intensive.

There may be many possible algorithms for context extraction and recognition that satisfy these three requirements. For illustration purposes, we next provide a description of one such algorithm. It is fully automatic and uses the Internet as a knowledge base to extract multiple contexts. The use of the Internet provides a language-independent mechanism and also avoids biases by applying query expansion to the original text, as provided by the organization expert. This algorithm was adapted from [30] and is currently part of the QUALEG solution.

The use of the Internet as a context database instead of a precalculated frequencies base [6] has several advantages. The use of the Internet does not require the constant updating and maintenance of a database, while the precalculated frequencies base requires the user to work in a limited predefined knowledge domain. Also, the Internet can serve as an unlimited knowledge domain that is continuously being updated. Last but not least, the multilingual nature of the Internet makes it a perfect infrastructure for the proposed method.

The success of the algorithm depends, to a great extent, on the number of documents retrieved from the Internet. With a greater number of relevant documents, less preprocessing (using methods such as Natural Language Processing) is needed in the data collection phase.

4.2.1. A context recognition algorithm

Let $D = \{P_1, P_2, \ldots, P_m\}$ be a set of textual propositions representing a document, where for all P_i there exists a

collection of descriptor sets forming the context $C_i = \{\langle c_{i1}, w_{i1}\rangle, \ldots, \langle c_{in}, w_{in}\rangle\}$ so that ist(C_i, P_i) is satisfied. McCarthy [19] defines a relation ist(C,P), asserting that a proposition P is true in a context C. The granularity of the textual propositions varies, based on the case at hand, and may be a single sentence, a single paragraph, a statement made by a single participant (in a chat discussion or a Shakespearian play), *etc.* The context recognition algorithm identifies the outer context C defined by

$$\text{ist}\left(C, \bigcap_{i=1}^{m} \text{ist}(C_i, P_i)\right).$$

The algorithm input is defined as a set of textual propositions representing a document. Each textual proposition is sent to a Web search engine. The set of descriptors is extracted by clustering the Web pages search results. The number of textual propositions that extract the same descriptor identifies the number of references to the descriptor in the text. Similarly, the number of Web pages that identify the same descriptor represents the number of references in Internet documents. A high ranking in only one metric does not necessarily indicate the importance of the context: for example, high ranking in only Internet references may mean that it is an important topic but might not be relevant to the document. To combine both metrics the two values are weighted to contribute equally to final weight value.

The context recognition algorithm consists of the following major phases: collecting data, selecting contexts for each text, ranking the contexts, and declaring the current contexts. The phase of data collection includes parsing the text and checking it against a stop-list. To improve this process, text can be checked against a domain-specific dictionary. The result is a list of keywords obtained from the text. The selection of the current context is based on searching the Internet for relevant documents according to these keywords and on clustering the results into possible contexts. The output of the ranking stage is the current context or a set of highest ranking contexts. The set of preliminary contexts that has the top number of references, both in number of Internet pages and in number of appearances in all the texts, is declared to be the current context and the weight is defined by integrating the value of references and appearances. An example of identifying the contexts that receive both high number of references and high number of appearances is illustrated in Fig. 2.

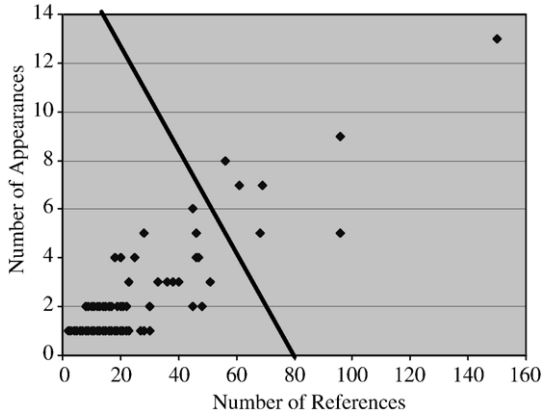

Fig. 2. Identifying the current contexts.

4.3. Querying the multilingual information system

We now turn our attention to querying, one of the main usages of knowledge management of information systems. When a user submits a query to the information system, it can be classified using the ontology to a specific concept, based on context comparison. As a concrete example, consider a new document that is submitted to the information system. We can extract a context from this document and then compare it to the existing contexts associated with ontology concepts. Matching contexts may lead to tasks such as document classification, email routing, workflow activity processing, *etc.*

Automatic context extraction is an uncertain process, subject to noise that exists in the documents at hand. Different context extraction algorithms may yield varying levels of uncertainty. In any case, it may be too restrictive to adhere to a strict approach according to which a context can be matched to an ontology concept only if it completely matches the concept's context. In context extraction, a generated false negative context can be, for example, Music, which is not represented in the Theater Festival concept but is required. Conversely, a context such as Art can also provide false positive identification of related documents, since not all of them are related to Theater.

The descriptors c_i and their respective weight w_i are extracted by the algorithm described in Section 4.2.1. In this algorithm, the weight represents the number of appearances in the text and the number of references in the Internet to the context.

The selection of the context is based on a search through the Internet for all relevant documents according to the text from the documents. The retrieved documents are clustered into possible descriptors. For each descriptor, we measure how many times it is referred to in the text and how many Internet pages refer to it. For example, Music might not appear at all in the document, but the descriptor based on clustered Internet pages could refer to it 2 times in the text and 235 Internet pages might be referring to it. The

following formula is used based on [30]. The descriptors that receive the highest ranking form the context. The WeightedValue forms the w_i weight previously described. The weight is calculated according to the following steps. Find the difference between each value of the number of references and its nearest lower value neighbor, defined as Current References Difference Value (CRDV). Find the difference between each value of the number of appearances and its nearest lower value neighbor, defined as Current Appearances Difference Value (CADV). The weight of the number of appearances in the text and the number of references in the Internet is calculated according to the following formula:

MVR = Maximum Value of References
MVA = Maximum Value of Appearances

$$\text{Weighted Value} = \sqrt{\left(\frac{2*\text{CADV}*\text{MVR}}{3*\text{MVA}}\right)^2 + (\text{CRDV})^2} \quad (1)$$

To compare contexts, we first define distance between two descriptors c_i and c_j with their associated weights w_i and w_i as follows:

$$d(c_i, c_j) = \begin{cases} |w_i - w_j| & c_i = c_j \\ max(w_i, w_j) & c_i \neq c_j \end{cases} \quad (2)$$

This distance function assigns greater importance to descriptors with larger weights, assuming that weights reflect the importance of a descriptor within a context. According to Eq. (2), if $c_i = c_j$, then $d(c_i = c_j)$ is the absolute difference between the weights. In other words, the equation measures the distance of the weights between identical contexts. In all other cases, it takes the maximal weight as the distance between them, which allows a lower value to be given only to similar context when calculating distance. To define the best ranking concept in comparison with a given context we use Hausdorff metric, as follows. Let A and B be two contexts and α and b be descriptors in A and B, respectively. Then,

$$d(a, B) = \inf\{d(a, b) | b \in B\} \quad (3)$$

$$d(A, B) = \max\{\sup\{d(a, B) | a \in A\}, \sup\{d(b, A) | b \in B\}\} \quad (4)$$

Eq. (3) provides the value of minimal distance of an element from all elements in a set. Eq. (4) identifies the furthest elements when comparing both descriptor sets.

To summarize, the mapping process introduced above is language independent. Therefore, relevant ontology concepts can be identified as long as the ontology context is sufficiently similar to the matched context. Such a context mapping process enjoys all the benefits mentioned above. Therefore, it allows variations of emphases that stem from local interpretation and lingual differences. In addition, its support of less-than-perfect mappings compensates for differences of terminology while ontologies alone often lack such flexibility. To understand why, one should recall that the ontology is designed with the assistance of an organization expert. However, the new arriving documents can come from laymen (*e.g.*, an email that arrives from a citizen), using a different terminology. It is hard, in particular for an organization expert, to anticipate such variations and design them as part of the ontology.

4.3.1. Examples

We now present three examples to illustrate possible usage of knowledge management with the proposed model in querying a multilingual information system. The three examples are taken from the eGovernment domain. The first example involves the routing of input, such as an incoming email, to the appropriate place in the organization. Given a distance threshold, t_1, any ontology concept whose context matches an automatically generated context from an email and its distance is lower than the threshold ($d(A,B) < t_1$) will be considered relevant. If such a context has an associated email address, the email will be routed to it. An overlap between contexts belonging to different concepts is possible, similar to dynamic taxonomies [27].

The second example involves opinion analysis. A relevant set of ontology concepts is identified, as in the case of email routing. The ontology also contains contexts defining various opinions. Such contexts may be globally defined (for the whole ontology) or specific to some concepts. Opinion contexts can be defined in multiple languages. In QUALEG, positive and negative words were taken from the WebUse Scientific Research On the Internet from the University of Maryland (http://www.webuse.umd.edu:9090/tags/). These words were translated using an online dictionary to the German language. The relative distances of the different opinions of a matching concept are evaluated. If the difference in distance is too close to call (given an additional threshold t_2), the system refrains from providing an opinion. Otherwise, the email is marked with the opinion with minimal distance.

The determination of public agenda is a third task the system can support. If all ontology concepts (of the n relevant concepts) do not exceed the threshold $d(A,B) \geq t_1$, then the email is considered to be part of a new topic on the

public agenda and is added to other emails under this concept. Periodically, such emails are clustered and provided to decision makers to determine the addition of new ontology concepts.

Fig. 1 presents an example of a multilingual ontology. Each concept is represented by a node with multiple contexts for each language. It contains three ontology concepts, namely Citizen, Public Service, and Activity of Public Service. Each ontology concept in English is translated to Polish, French, and German. Next, there are ontology concepts that are relevant only to the local government of Tarnow and therefore they appear only in English and in Polish. Each local government can extend the ontology concepts to include ones that interest it alone and can decide to use existing ontology concepts simply by adding the translation to the local language.

Consider the following example of an email in Polish, describing an email received by local government of Tarnow on the topic taxes on immovable assets:

Subject: podatek od nieruchomości
Szanowni Państwo,
Zwracam się z prośbą o przesłanie wysokości stawek opłat za podatek od nieruchomości dla osób prawnych w Państwa mieście obowiązujące w latach 2000–2004?
Z poważaniem
Edyta

The context as extracted by the context recognition algorithm include: {⟨Podatek, 59⟩, ⟨nieruchomości, 43⟩, ⟨Polska, 26⟩, ⟨Strona, 21⟩}. The first two, translated to value added tax (v.a.t.) and immovables, are very relevant to the topic of the email. The other two contexts, Polska, which is Poland, and Strona, which has multiple meanings such as a party or side, have less relevance to the scenario described in the text.

When mapping the contexts to the ontology, the context of nieruchomości can be identified in the list. Therefore, this document can be mapped to the topic of Local Spatial Management Strategic Plan, which can now be accessed by both English or Polish queries.

The following is an example of a German email:

Strassentheater und das kostenlos und auf hohem iveau. Ein Aushängeschild für Saarbrücken und ein leuchtendes Beispiel für junges wildes Theater, abseits des Mainstreams.
Toll.

The context recognition algorithm identified the following context, described in detail in Table 1 with the actual values assigned by the algorithm. The results are represented by the context on the left side. There are two possible main ontology concepts, Perspectives du Theatre and Long Day School, to which each data item can belong. The Perspective du Theatre concept encompasses six other subconcepts, which include Rahmenprogramm, Organisation, Spielplan, Veranstalter, Besucher, and Informationen, when each context can belong to one or more concepts. Note that the context is extracted not only in German but also in French. The highest ranking was received by the Perspective du Theatre concept. Note also that some of the contexts are not mapped to any concept. Therefore, when we examine which of the subconcepts related to Perspective du Theatre are relevant to the text, we can classify the document as belonging to the Rahmenprogramm (master program) in Perspective du Theatre which received the total highest score out of all subconcepts.

5. Experiences with QUALEG

QUALEG has pilots in France, Poland, and Germany and thus currently focuses on four languages, three of which are French, Polish, and German. English is also used as a common international representation language. To maintain uniformity and avoid repetitive translations, QUALEG processes the information from the input, such as debates and emails, in the local languages.

Table 1
Algorithm sample results

Context Descriptor	Perspectives du Theatre Concept Relevance	Long Day School Concept Relevance
Art	10 Rahmenprogramm, Organisation, Spielplan, Veranstalter	1
Abseits des Mainstreams	0	0
Oscar Wildes	0	0
Jenseits des Mainstreams	0	0
Movie	0	0
Firma	1 Besucher	0
Saar	2	0
Download	0	0
Programm	3 Rahmenprogramm, Informationen	0

For the deployment in QUALEG the first step included starting with an existing ontology and expanding it for the specific project needs. The ontology used was a local government ontology developed for TerreGov, a different EU project. In the collection step, local government representatives from each of the pilots supplied organizational documents that describe each concept in the ontology collected from previous years. The extraction step created a context for each concept in the ontology using the algorithm described in Section 4. The last step involved adding the new contexts to their relevant concepts and storing them. These contexts are monitored according to the system performance and can be updated when needed.

The system is built to support multilingual ontology management. The system allows an ontology search to be performed, retrieving documents that relate to a specific ontology concept. The mapping of the documents to the ontology concepts is performed using the context recognition algorithm implemented in the Knowledge Extraction module.

Section 5.1 presents our experiences with the German language. Section 5.2 discusses the advantages this technique has in some languages versus others and the extent of language independence of the model. We conclude (Section 5.3) with applications of the model of ontology and context in the field of opinion analysis.

5.1. The German Perspectives du Theatre Festival

Our first experience was with the Perspectives du Theatre Festival held during May every year in Saarbrücken, located at the French border of Germany. The festival includes contemporary French theatre, films, street events, music, *etc*. Our challenge was to analyze the material and provide a useful set of classifications so that the materials could be rapidly understood and routed to the appropriate civil servants.

The data we received included daily communications (in German) about this event, consisting of 104 different emails, primarily emails from citizens to the city hall and press releases and announcements from the city outward. The festival is an annual event and we were given data from 2004 and 2005.

The goal of the topic classification experiment was to identify the topic of the email according to a predefined list of ontology concepts as supplied by Saarbrücken for organizing cultural events. The predefined concepts of the emails supplied by Saarbrücken were: Organisation, Veranstalter, Finanzen, Besucher, Informationen, Rahmenprogramm, Spielplan, Other. Each topic, an ontology concept, was accompanied by a set of contexts that describe it.

To examine the proposed model we used a single ontology and two different methods to define and extract contexts. One method was the one described in Section 4.2.1. This method used the technique of mapping contexts to ontology concepts, as detailed in Section 4.3. Each set of words was mapped from the email to the Internet, extracting a set of best ranked textual contexts that define the document. These contexts were searched against the list of contexts describing each concept in the ontology. The other method was based on conventional Natural Language Processing techniques, enhanced by a language domain expert to build a set of rules for identifying relevant words and grammar relevant to the German language. This technique was based on a per sentence analysis. For each sentence a classifier, automatically trained on keywords and morphological variants (based on the initial list of topics from Saarbrücken), was used. Each sentence in the email was searched against the list of keywords and morphological variants. The two techniques are very different. The former is language independent, making it more suitable for multilingual environments at the possible cost of lacking language-specific analysis tools, used by the latter.

The experiment included classifying the incoming data according to the concepts described above. We compared the recall and precision of the proposed model to the Natural Language Processing technique. The input to both methods was identical. The context extraction technique were different. The output of both methods was a best matching concept. The data was also analyzed by two Natural Language experts and a local government representative from Saarbrücken to supply the "golden standard."

For both methods the input was parsed at the granularity of sentences. Our Internet search based technique parsed long sentences according to the maximum number of words that could be used in a search engine. The Natural Language Processing preprocessing included a Tokenizer, a tool for breaking up compound nouns, and a German Demorpher (Morphy engine), downloaded from the University of Stuttgart (http://www.lezius.de/wolfgang/morphy/). The Demorpher removes case markings, tense markings, *etc*.

Two different experiments were performed. The first experiment was to analyze our model based on the German data. The knowledge extraction component achieved a Precision of 85.37%, Recall of 84.34%, and total F-Score of 84.85%. This is based on the comparison of the results of the Context Recognition/Knowledge Extraction component to the human judgements. The German input data was classified by two German Language experts and by Saarbrücken local government civil servant employees.

The second experiment analyzed the performance of our method implemented in the Knowledge Extraction compared to the NLP technique. In this experiment a subset of 72 different emails representing data from a single year was used for comparison. The number of emails in the second experiment is a subset of the total 104 emails since only opinion-related emails were used in this experiment as opposed to official announcements messages that were sent at the time. The results show that method proposed in this work achieved the *F*-Score of 81% while the Natural Language Processing technique achieved the *F*-Score of 78% where the precision and recall were weighted equally. The results show that the proposed ontology-based multilingual model of contexts and ontology achieved better results.

These results show the promising ability of the proposed model to provide language-independent support to local government decision making.

5.2. Extent of language independence

Having shown a proof-of-concept of the proposed model applicability, we next discuss the extent to which language-independent algorithms, such as the one proposed in Section 4.2.1, can be used in a multilingual setting. The proposed algorithm compensates for the lack of language-specific rules by using a vast database, such as the Internet, to gain better statistical knowledge in identifying contexts. However, there is a skewed distribution of languages over the Internet. Xu [34] estimated that 71% of the pages (453 million out of 634 million Web pages indexed by the Excite search engine at that time) were written in English, followed by Japanese (6.8%), German (5.1%), French (1.8%), Chinese (1.5%), Spanish (1.1%), Italian (0.9%), and Swedish (0.7%). Earlier experiments [30] show over 90% recall of the proposed algorithm for the English language. Our experience with the German language also suggests reasonable performance, with an F-Score of above 80%. Would such a success rate be retained with languages whose relative presence on the Web is much lower? Some researchers argue that one hundred million words is a large enough corpus for many empirical strategies for learning about language, either for linguists [2] and lexicographers [16] or for technologies that need quantitative information about the behavior of words as input (most notably parsers [17]).

To test the extent to which language-independent algorithms can be used, we tried a classification task with the Polish language. The relative presence of the Polish language on the Web is less than 0.5%, 10 times smaller than the German presence [16]. This means that there are a few million Polish Web pages, which according to previous research may not suffice for this task.

For the purpose of our experiment, we analyzed four ontology concepts, namely Przyroda, Transport, Zabytki and Zagospodarowanie, in the local government of Tarnow. A total of 30 documents were analyzed. For each concept, contexts were identified manually by Tarnow local government civil servant employees and each of the documents was classified using the algorithm in Section 4.2.1. Our initial experiment, which avoided the use of any language dependent tool, yielded poor results of less than 11% recall. Therefore, we applied a simple NLP mechanism. We used a synonym dictionary on the results of the context recognition algorithm. The Polish dictionary in Portal Wiedzy Tłumacz (http://portalwiedzy.onet.pl/tlumacz.html) provides multiple synonyms and demorphing for each word translated. The use of the dictionary for the identification of synonyms and demorphing increased the number of contexts and thus increased the chances that the words associated with the ontology concept would be identified. The use of these tools increased the recall to 97%.

Due to the small sample we had at our disposal, no conclusive conclusions can be given at this time, and more experiments are needed. Nevertheless, this experiment indicates that for languages with smaller presence on the Internet, the proposed algorithm needs to be enhanced with language-specific methods. We note that language dependent tools are available and can easily be combined in a multilingual system. Other methods require a more in-depth knowledge of a language and may be specifically tailored to a given domain. Using more of the former (as we did with the Polish language) and less of the latter improves the deployment of information systems across borders.

5.3. Multilinguality and opinion analysis

Another possible application of the multilingual model lies in the field of opinion analysis. Opinions can be viewed as perspectives expressed in the input information. Opinions can be included in the ontology as concepts, associated with sets of contexts that provide the local interpretation of each opinion.

In the QUALEG project the system was developed in two different parts, separating the task of knowledge extraction from that of opinion analysis. The main difference between the two parts of the system is that the knowledge extraction component avoids the language-specific implementation and bases its analysis techniques on the use of a large corpus of relevant documents taken from the Internet, while the opinion analysis component

uses techniques from IR and NLP to improve content understanding. As in the knowledge extraction, the results of the opinion analysis are mapped to concepts in the ontology, in this case, opinion concepts. These opinions fall into three categories of concepts — positive, negative, and neutral.

The experiment included 72 emails in German to analyze the opinion analysis component. For the use of opinion analysis a set of opinion words were analyzed in English and machine translated to German. These opinion words are associated with the three opinion concepts. Two possibilities were examined: first, to translate the emails into English and then analyze the texts for opinion, and second, to translate opinion words. The latter alternative was found to achieve better results, resulting in slightly better accuracy of 60% versus 59% of the first option. These results tested the identification of the correct opinion words of each sentence. The results can be explained as slightly better since fewer words are translated. These translations were based on the European Parliament corpus using GIZA ++, followed by the alignment intersection heuristic [22].

The final results included adding additional NLP information in German. For example, in German the last opinion word in a sentence overrules previous opinion words. In addition, the opinion of each sentence was analyzed separately and summed to identify the opinion of the whole email. The results of the opinion analysis reached a Precision of 78.95%, Recall of 69.23%, and *F*-Score of 73.77%. The results indicate that expanding the context and ontology model to perform opinion analysis is feasible, albeit at a lower accuracy.

5.4. Experiments

To analyze the impact of our model for the support of multilinguality in information systems we experimented with data from news RSS. RSS is a format for distributing and gathering content from sources across the Web, including newspapers, magazines, and blogs. Web publishers use RSS to easily create and distribute news feeds that include links, headlines, and summaries. As a result, RSS serves as a useful platform for comparing news items in different languages supplied by different sources. We start with a description of the real-world data trace and experiment set-up, followed by a description of our experiments and an empirical analysis of the results.

5.4.1. Data sets and metrics

The RSS news data traces come from BBC — British English news (http://news.bbc.co.uk/1/hi/help/3223484.stm), CNN — American English news (http://edition.cnn.com/services/rss), Stern — German news (http://www.stern.de/sonst/?id=517321), and Le Figaro — French news (http://www.lefigaro.fr/rss/). We use the first two as two sources with local interpretations of topics.

The list of topics we selected from each news data trace is displayed in Fig. 3, where selected topics are circled. Topics are taken from four categories, namely Politics, World, Science, and Technology. Topics may vary slightly among different RSS sites. Some sites unify topics, *e.g.*, Science and Medicine in Le Figaro. In what follows, a concept is either a topic or a category, based on the experiment.

In these data traces data are partitioned to topics with no ontological relationships. The experiments focus on the concepts/contexts relationships, for which these data sets serve adequately. Research and experiments on ontological relationships using contexts are reported in [32].

The RSS trace was collected during August–September 2006. The news topics in each data set include between 33 and 641 data, where a datum is an RSS news header or a news descriptor. There was a total of 1778 data items used in the experiments. Table 2 describes the RSS news data sets. The table summarizes the number of news data for each news data set (size), the number of categories, the minimum and maximum number of data per concept, and each concept size. Concept size represents the total number of data in up to the four data sets.

We generated a context for each concept using the algorithm described is Section 4. This context is referred to as context* and the data that was used for this context generation is referred to hereafter as the context* data. The number of data items that were used for generating context* data was set to 10 and to 51 data items, based on empirical analysis performed in [32]. The context* data are selected randomly from the data items associated with a concept. A varying number of concepts was used, ranging from 1 to 13 concepts depending on the experiment. We also experimented using only four concepts, representing the four categories. In this case, the data for generating context* were chosen randomly from the multilingual data set. For context extraction we use the algorithm, adapted from [30]. Our aim is to test the impact of multilinguality on the performance of our model, rather than to test the text classification abilities of this or that algorithm. Nevertheless, this algorithm is known to have generated reasonable contexts in the past (see experiments in [30]).

As a measure of evaluation we use recall and precision metrics. The recall of a concept is defined as the ratio of data items which share at least one descriptor with the descriptors of context* and the number of the data items which belong to the concept. A high recall measure means that the algorithm was able to classify correctly a good

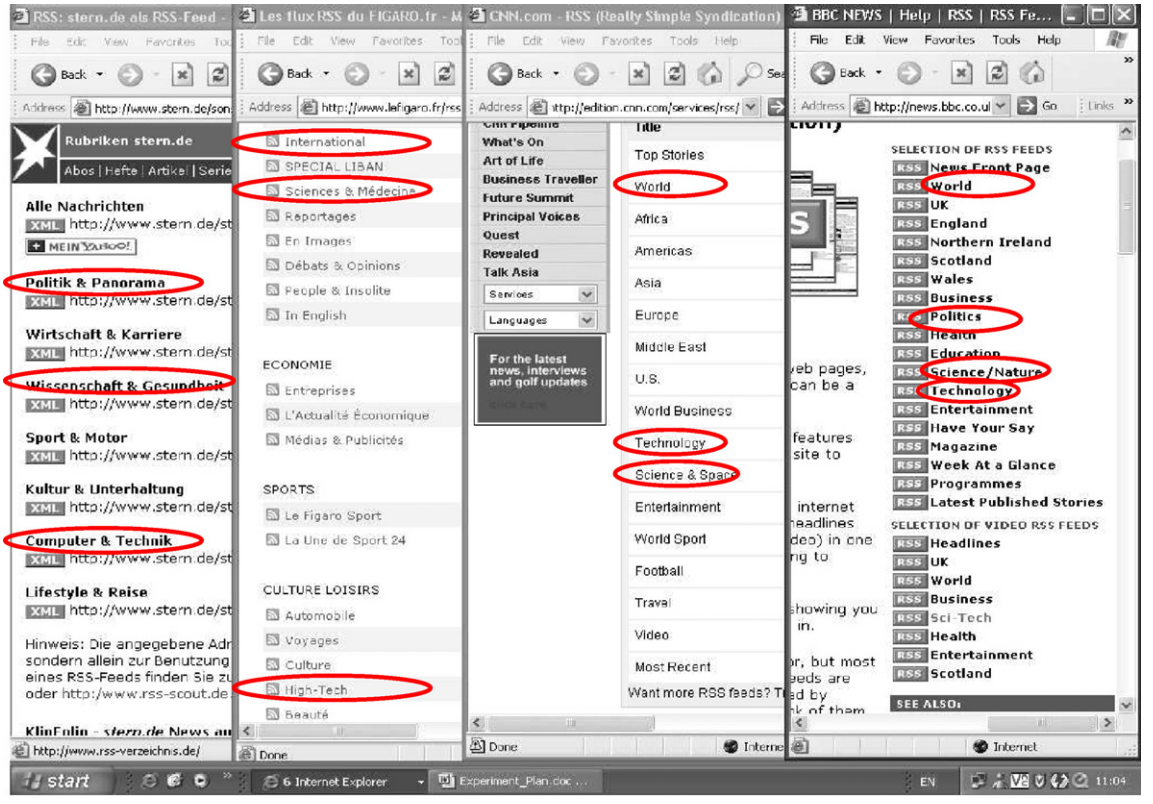

Fig. 3. Experiment outline.

portion of the data item set, minimizing false negative. Precision is defined to be the ratio of the number of true positive identifications and the number of data items associated with a concept. We measure precision with respect to the original classification of data items to topics/categories as given in the data traces. It is worth noting that in most of the experiments the algorithm classifies a datum to a *single* concept, the one whose contexts share the highest number of descriptors with its context, thus setting a lower bound on the algorithm performance.

5.4.2. Experiment results

We are now ready to discuss our experiments in details. The overall purpose of these experiments is to test the impact of multilinguality on our model. We start with evaluating the model ability to correctly classify documents to concepts in various languages. Then, we analyze the impact of having a multilingual corpus on the precision of classification. Finally, we analyze the impact of local interpretations, showing the need to compensate ontologies for under-specification.

5.4.2.1. Single class classification. In the first experiment, we evaluated the impact of multilinguality on classification recall. In each experiment we selected a single concept and generated a context using context* data. For each of the remaining data in this category a context was generated and compared with the context*. For each context* data size we repeated the experiment 4 times, each time choosing randomly the context* data. To evaluate the impact of multilinguality on a single class classification, we also generated four multilingual data sets, each representing a single category and repeated the experiment with these data sets.

Table 2
RSS data set statistics

Data set	BBC	CNN	Stern	Le Figaro
Size	1007	257	246	268
Categories	4	3	3	3
Minimum data per category	68	42	33	48
Maximum data per category	641	159	157	156

Data set	Politics	Science	Technology	World
Size	358	266	198	956

A graphic illustration of our results is given in Fig. 4, where the top part provides an overview of the results and the bottom part focuses on the recall range of 90%–100%. The *y*-axis displays the average recall over the four experiments of each data set. Each group represents a different category. The last bar of each group shows the average recall of the multilingual data set. The right-most group provides an average recall over all data sets. In this experiment each context, besides context*, is limited to 10 descriptors.

A per concept analysis shows an average recall ranging from 91.67% to 100% in the news RSS data set, when context* is defined using up to 45 descriptor sets. The impact of multilinguality is seen by comparing the average recall of the "joint" bar vs. the individual topic results. On average, the use of multilingual corpus results in a minor reduction of less than 2% in recall (from about 98.35% to 96.58%). A category-based analysis shows that in one case the joint results are lower than individual data sets average recall, in two cases they fall in between other results, and in one case they reach the top results. All-in-all, the impact of using a multilingual corpus is negligible.

5.4.2.2. Multiple class classification. Next, we analyze the impact of multilinguality on classification precision. For this set of experiments we enforced a rigid classification scheme, in which each document is classified to a single concept. It is worth noting that not all applications follow such a strict approach. In QUALEG, for example, email routing requires emails to be routed to all concepts whose context is sufficiently similar to the email's context.

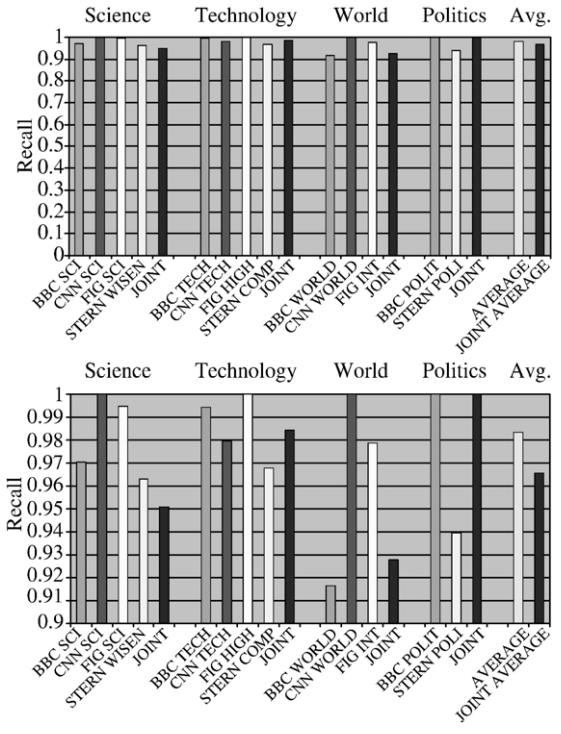

Fig. 4. RSS recall.

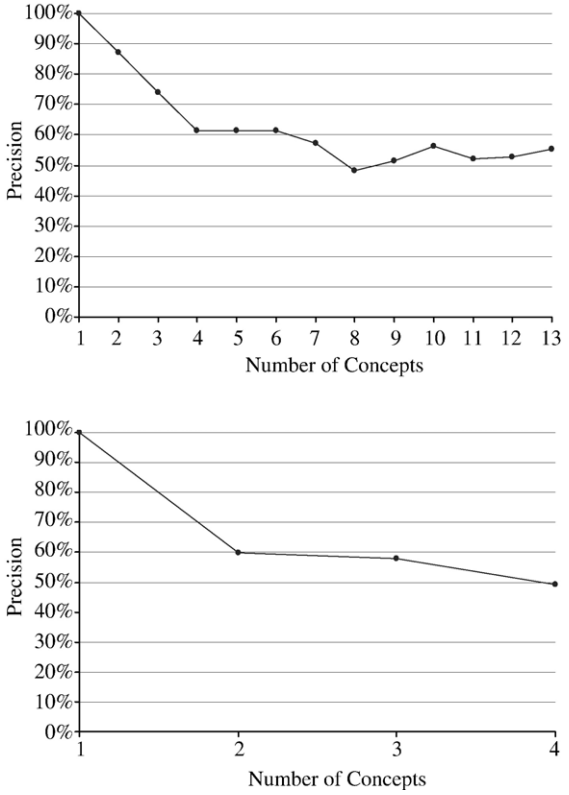

Fig. 5. RSS precision.

The experiment results are summarized in Fig. 5. The horizontal axis displays the number of participating concepts and the vertical axis presents the classification precision. Fig. 5(top) analyzes the 13 concepts separately. In Fig. 5(bottom) we present the results for four multilingual corpora, as described in the previous experiment. In this experiment each context was limited to 10 descriptors and context* is defined using up to 45 descriptor sets.

As the number of concepts increases, precision declines, stabilizing at a level of 50%–60% after 7 concepts. To analyze the impact of the multilinguality we compare the results displayed in the two graphs. The precision for all 13 concepts reaches 55.17% while the use of multilingual corpora reduces the precision by about 6% to 49.06%. Even when ignoring the data corpus sizes, and comparing to the performance of classifying to four classes in Fig. 5 (bottom), one observes a reduction of 10.84%, from 59.9% to 49.06%. These results indicate that our model suffers a minor reduction in performance with the introduction of multilinguality.

Table 3
Average recall for BBC and CNN

Recall		Training		
		BBC	CNN	BBC+CNN
Data	BBC	96.06%	75.31%	94.58%
	CNN	97.68%	99.32%	99.90%

5.4.2.3. Analysis of local interpretation. This experiment analyzes the performance of our model, when presented with local interpretation of concepts in the same language. For each context* data size we repeated the experiment 4 times, each time choosing randomly the context* data. We compared three similar concepts of CNN and BBC, namely World, Technology, and Science, as well as three similar concepts of BBC and Stern, Politics (Politik & Panorama), Technology (Computer & Technik), and Science (Wissenschaft & Gesundhelt), as a control set. We used one data set for training and then used the other data sets as test data, interchanging the roles of news agencies. We also analyzed the results of uniting two data sets into one when selecting a similar size of data from each data set concept. The united data set was used as a training set and each of the two original data sets was used as test sets.

Average recall results over the three topics, tested separately, are displayed in Table 3. When using one data set for training and another for testing, recall is lower (75.31% and 97.68%) than when training data and test data are taken from the same data set (96.06% and 99.32%, respectively). The impact of our model is evident in the right-most column of Table 3. Using data from both BBC and CNN for training improves recall (94.58% vs. 75.31% and 99.90% vs. 97.68%) and is similar to that of using a homogeneous data set for both training and testing (94.58% vs. 96.06% and 99.90% vs. 99.32%). For CNN data, the use of both data sets for training slightly increases recall (from 99.32% to 99.90%), a result that may be attributed to statistical variation.

Average precision results for CNN and BBC are displayed in Table 4 and CNN and Stern results are displayed in Table 5. Precision drops when using training from a different data set, *e.g.*, training with CNN for BBC data reduces precision from 68.15% to 47.31%. These results serve as an empirical justification of our claim that even within the same language, local interpretation has a major impact, and therefore even without the language barrier, ontologies quickly become under-specified. Another observation is that the use of a combined data set for training does not improve precision. This means that whenever an information system is deployed for local use only, it is better to avoid using contexts from other deployments. As a control for our experiments, we tested precision also by comparing BBC with Stern, generating multilingual contexts. We observed the same phenomenon in a multilingual setting as well, although precision is better here. One explanation for the improved precision may be the higher word variation in different languages.

Table 4
Average precision BBC and CNN

Precision		Training		
		BBC	CNN	+CNN
Data	BBC	68.15%	47.31%	37.03%
	CNN	47.09%	59.60%	35.54%

Table 5
Average precision BBC and Stern

Precision		Training		
		BBC	STERN	BBC+STERN
Data	BBC	73.64%	73.53%	40.12%
	STERN	60.31%	72.08%	44.66%

6. Conclusion

In this work we proposed a knowledge management model for the support of multilingual applications. The model is based on a global ontology, manually designed for a specific domain, and local contexts, associated with ontology concepts. The combination of ontologies and contexts lends itself well to multilingual applications in which a single ontology fails to capture all nuances that stem from language and cultural differences. The model was presented both in technical terms and via an example from the eGovernment domain. The model properties were discussed and some experiences with a specific eGovernment application, QUALEG, were described and analyzed.

The single ontology system with associated concepts in multiple languages proposed here provides a framework that is both versatile and flexible. The system functions simultaneously in multiple languages, is low-maintenance, and is easily extended in and adapted to different languages. The model captures cultural as well as lingual differences using contexts, thus allowing easy customization across cultures and languages.

Future directions of research include identifying methods for allowing real-time interaction between local government representatives and citizens through the use of multilingual ontologies. Another direction is identifying

the ability to recommend a policy to the local government, based on the information in the ontology, and automatic translation of words that define the ontology based on their context. In addition, future research can examine the performance of the system when implemented on languages with different character sets, such as Chinese and Arabic.

Acknowledgments

The work was partially supported by two European Commission 6th Framework IST projects, TerreGov (http://www.terregov.eupm.net) and QUALEG (http://www.qualeg.eupm.net), and the Fund for the Promotion of Research at the Technion.

References

[1] S. Aytac, Multilingual information retrieval on the internet: a case study of Turkish users, International Information & Library Review 37 (2005) 275–284.
[2] C.F. Baker, C.F. Fillmore, J.B. Lowe, The Berkeley framenet project, In Proceedings of COLING-ACL, 1998, pp. 86–90.
[3] J. Bar-Ilan, T. Gutman, How do search engines handle non-English queries? A case study, In Proceedings of the twelfth international World Wide Web conference, 2003.
[4] M. Bunge, Treatise on basic philosophy, The Furniture of the World, Ontology I, vol. 3, D. Reidel Publishing Co., Inc., New York, NY, 1977.
[5] M. Bunge, Treatise on basic philosophy, A World of Systems, Ontology II, vol. 4, D. Reidel Publishing Co., Inc., New York, NY, 1979.
[6] R. Chau, C. Yeh, A multilingual text mining approach to web cross-lingual text retrieval, Knowledge-Based Systems 17 (2001) 219–227.
[7] F.M. Donini, M. Lenzerini, D. Nardi, A. Schaerf, Reasoning in description logic, in: G. Brewka (Ed.), Principles on Knowledge Representation, Studies in Logic, Languages and Information, CSLI Publications, 1996, pp. 193–238.
[8] A. Gal, A. Anaby-Tavor, A. Trombetta, D. Montesi, A framework for modeling and evaluating automatic semantic reconciliation, VLDB Journal 14 (1) (2005) 50–67.
[9] A. Gal, A. Segev, Putting things in context: dynamic eGovernment re-engineering using ontologies and context, In Proceedings of the 2006 WWW Workshop on E-Government: Barriers and Opportunities, 2006.
[10] T.R. Gruber, A translation approach to portable ontologies, Knowledge Acquisition 5 (2) (1993).
[11] A. Hotho, S. Staab, A. Maedche, Ontology-based text clustering, In Proceedings of the IJCAI— 2001 Workshop Text Learning: Beyond Supervision, 2001, p. 29.
[12] E. Hovy, N. Ide, R. Frederking, J. Mariani, A. Zampolli, Multilingual information management: current levels and future abilities, Linguistica Computazionale XIV–XV (2001).
[13] J. Hutchins. Current commercial machine translation systems and computer-based translation tools: System types and their uses. International Journal of Translation, 2005. to appear.
[14] J. Kahng, D. McLeod, Dynamic classificational ontologies for discovery in cooperative federated databases, In Proceedings of the First IFCIS International Conference on Cooperative Information Systems (CoopIS'96), June 1996, pp. 26–35, Brussels, Belgium.
[15] M. Kifer, G. Lausen, J. Wu, Logical foundation of object-oriented and frame-based languages, Journal of the ACM 42 (1995).
[16] A. Kilgarriff, G. Grefenstette, Introduction to the special issue on the web as corpus, Computational Linguistics 29 (3) (2003).
[17] A. Korhonen, Using semantically motivated estimates to help subcategorization acquisition, In Proceedings of the Joint Conference on Empirical Methods in NLP and Very Large Corpora, 2000, pp. 216–223.
[18] T. Liu, Z. Chen, B. Zhang, W.-Y. Ma, G. Wu, Improving text classification using local latent semantic indexing, In ICDM, 2004, pp. 162–169.
[19] J. McCarthy, Notes on formalizing context, In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, 1993.
[20] C. Mooers, Encyclopedia of library and information science, vol. 7, Marcel Dekker, 1972, pp. 31–45, chapter Descriptors.
[21] H. Moukdad, Lost in cyberspace: how do search engines handle Arabic queries? In access to information: Technologies, skills, and socio-political context. In Proceedings of the 32nd annual conference of the Canadian Association for Information Science, 2004.
[22] F.J. Och, H. Ney, Improved statistical alignment models, 38th Annual Meeting of the Association for Computational Linguistics, 2000, pp. 440–447.
[23] T. Ong, H. Chen, W. Sung, B. Zhu, Newsmap: a knowledge map for online news, Decision Support Systems 39 (2005) 583–597.
[24] N. Papadakis, A. Litke, D. Skoutas, and T. Varvarigou. Memphis: A mobile agent-based system for enabling acquisition of multilingual content and providing flexible format internet premium services. Journal of Systems Architecture, 2005. to appear.
[25] H. Putnam (Ed.), Reason, Truth, and History, Cambridge University Press, 1981.
[26] S. Russell, P. Norving, Artificial Intelligence: A Modern Approach, 2nd edition, Prentice Hall, Upper Saddle River, New Jersey, 2003.
[27] G. Sacco, Dynamic taxonomies: a model for large information bases, IEEE Transactions on Knowledge and Data Engineering 12 (2) (2000) 468–479.
[28] H. Saggion, H. Cunningham, K. Bontcheva, D. Maynard, O. Hamza, Y. Wilks, Multimedia indexing through multi-source and multi-language information extraction: the MUMIS project, Data and Knowledge Engineering 48 (2004) 247–264.
[29] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys 34 (1) (March 2002) 1–47.
[30] A. Segev, Identifying the multiple contexts of a situation, Proceedings of IJCAI—Workshop Modeling and Retrieval of Context (MRC2005), 2005.
[31] A. Segev, A. Gal, Ontology verification using contexts, In Proceedings of ECAI—Workshop on Contexts and Ontologies: Theory, Practice and Applications, 2006, p. 30.
[32] A. Segev, A. Gal, Putting things in context: a topological approach to mapping contexts to ontologies, Journal on Data Semantics IX (2007) 113–140.
[33] P. Vossen, Eurowordnet general document, LE2-4003 LE4-8328, EuroWordNet, 1999.
[34] J.L. Xu, Multilingual search on the World Wide Web, In Proceedings of the Hawaii International Conference on System Science (HICSS-33), 2000.

Avigdor Gal received the DSc degree in temporal active databases from the Technion — Israel Institute of Technology in 1995. He is an Associate Professor at the Faculty of Industrial Engineering and Management, Technion. He has published more than 70 papers in journals, books, and conferences on data integration, temporal databases, information systems architectures, and active databases. He is a member of the steering committee of IFCIS, a member of IFIP WG 2.6, and a recipient of the IBM Faculty Award for 2002–2004. He is a member of the ACM and a senior member of the IEEE and the IEEE Computer Society.

Aviv Segev is an Assistant Professor at the College of Commerce, National Chengchi University. Previously, he was a postdoc at the Faculty of Industrial Engineering & Management at the Technion. In 2004 he received his Ph.D. from Tel-Aviv University in management information systems in the field of context recognition. During his studies, Aviv received the Vatat Scholarship for Excelling Doctorate Students in Elite Technology and the Adams Institute Scholarship Award. His current research includes classifying information and opinions of textual data, mapping of context to ontologies, and mapping of information. He has published a number of papers in scientific journals and conferences. Previously Aviv was a simulation project manager in the Israeli Aircraft Industry.