The Global Perpetual Dictionary of Everything

Helen Ashman, Associate Professor, School of Computer and Information Science, [HREF1], University of South Australia, [HREF2], GPO Box 2471 Adelaide, South Australia 5001. mailto:Helen.Ashman@unisa.edu.au

Dong Zhou, Student, School of Computer Science and Information Technology, [HREF3], University of Nottingham, [HREF4], Jubilee Campus, NG8 1BB, United Kingdom. mailto:dxz@cs.nott.ac.uk

James Goulding, Lecturer, School of Computer Science and Information Technology, [HREF3], University of Nottingham, [HREF4], Jubilee Campus, NG8 1BB, United Kingdom. mailto:jog@cs.nott.ac.uk

Timothy Brailsford, Lecturer, School of Computer Science and Information Technology, [HREF3], University of Nottingham, [HREF4], Jubilee Campus, NG8 1BB, United Kingdom. mailto:tjb@cs.nott.ac.uk

Mark Truran, Lecturer, School of Computer Science, [HREF5], University of Teesside, [HREF6], Tees Valley, TS1 3BA, United Kingdom. mailto:m.a.truran@tees.ac.uk

ABSTRACT

Knowledge organising systems aim to classify, store, organise and structure data for the purposes of retrieval. They usually rely on algorithms designed to automate these functions. However a largely untapped but highly accurate resource which can contribute enormously to these tasks is the mass of human intelligence which daily accesses that data. User consensus and judgement, or "co-active intelligence" is the emergent property arising from this mass human interaction with data, and there are now under development algorithms which exploit the consensus arising from this interaction. In particular, this consensus can be used to determine meaning, and can be used directly in search tools but can also contribute to the building of a complete, organic dictionary of almost anything that appears on the Web.

The primary advantage of co-active intelligence in the derivation of meaning is that it does not rely on increasingly complex sets of language-use rules, which must otherwise be elicited, represented and calculated. All such rules, while operating in co-active systems, are wholly implicit and are only ever activated by users applying their own judgement, thus never needing to be elicited. Co-active intelligence bypasses the explicit representation and automatic calculation of language-use rules, and acts only as a mediator between question and answer, leaving humans to answer the queries of human questioners.

INTRODUCTION

Web semantics are challenging to implement, requiring an infrastructure of ontology and interoperability and the creation of semantically-enriched data. This paper describes the co-active intelligence approach to semantic classification and association of Web resources that requires neither explicit ontology generation, nor the creation and maintenance of semantically enriched data.

The aim of this ongoing project, which we give the working name of the Global Perpetual Dictionary of Everything, is to generate a self-organising, self-maintaining, universal archive of meaning/aboutness of all addressable Web resources, text, images or otherwise, using the co-active intelligence (CI) principle which exploits user consensus to determine meanings or aboutness.

The resulting system will be a:

1. *global*: by default it applies to every addressable object on the Web, regardless of location. Even where "dark matter" Web objects are not seen by global search engines due to lack of ingoing links or other problems (Bailey et al. 2000), it is still possible for the principles to be used locally and generate "local" meanings, accessible locally, in the same way that search engines still function locally, even though the meanings are only suggestible to users who have access to those pages;
2. *perpetual*: it will be self-maintaining as it will constantly reassess the "centre of gravity" for a particular term as the user consensus shifts. By building it as a versioning system which archives previous meanings it is possible to preserve all meanings;
3. *dictionary*: we call it a "dictionary" although its real purpose is to collect and group Web

objects in order to exemplify meanings, rather than provide explicit definitions for terms. The meaning resulting from submitting the query is gleaned from the set of results returned from the query, whose sense is accurately confirmed by user consensus. It can of course also provide thesaurus and translational services;

4. *of everything*: every addressable object on the Web is incorporated, applying equally to multimedia objects such as images, and to text objects in other languages or belonging to localised or restricted glossaries.

While we call it a "dictionary" it is not identical to a print dictionary, since there will be not necessarily a definition returned after a query but rather a set of resources which are deemed by many other users to be "about" the same thing, i.e. to share the same or very similar meaning. Of course, print or online dictionary entries would form part of the resources, so a definition is frequently going to be part of the resources.

It operates on the principle of deriving meaning from a set of objects forming the results list of a query. This is indeed a dictionary in that a user seeking a meaning for a term will enter that term and receive one or more Web objects in a set of results, which may include documents, images or other multimedia, perhaps even formal definitions from a dictionary. While a comprehensive print dictionary offers a set of example usages of a term to supplement the given definition, the Perpetual Dictionary does the converse, by offering a set of examples with perhaps a formal definition to supplement. However in each case, the user requires an explanation of a term, and resources to help them comprehend that term are returned.

While the Perpetual Dictionary could be implemented directly on the back of a normal search engine, the accuracy of the co-active intelligence approach can greatly improve the relevance of the result set. Search engines or any information retrieval application suffer from various problems in relevance and precision of retrieved results, and "semantic drift" where elements of the language acquire new or altered meanings. Any formalised system which specifies rules for the application of meaning can only react to semantic drift after a critical mass of change has been perceived. In contrast, co-active intelligence is based on ongoing observation of users' activities, specifically, their selections from a search engine's results set - there is nothing "semantic" about a co-active intelligence system, meaning is only with its users. These observations and the conclusions drawn from them are constantly updated, so that the current usage(s) of a term is indicated almost as soon as it happens. Co-active intelligence uses no rules which specify meaning but instead assumes meaning is implicit in the consensus of previous users, as they select relevant documents from a result set in answer to their query term. In other words, a formal dictionary specifies a definition for a term, while the Perpetual Dictionary offers resources from which the meaning is emergent.

The Perpetual Dictionary implements a co-active intelligence search algorithm, operating between the actions of a normal search engine and the user. Data previously collected from users is used to classify the documents selected by the user from the search engine's result set as being relevant to the query term.

In the following sections we first outline the principle of co-active intelligence which underlies the Perpetual Dictionary, then discuss how it can be exploited to create it, and finally turn to how it can be extended to provide translational and thesaurus services.

CO-ACTIVE INTELLIGENCE

The judgement emergent from user consensus

The vast quantity of information and user interaction over the Web offers an opportunity to exploit human judgement as part of the classification processes in retrieval systems. Search engines are the focus for enormous numbers of transactions from users worldwide. Each query and, more importantly, the user's choices made from each set of query results, embodies the application of human judgement to a specific querying problem, with the most appropriate results generally being chosen earliest [1]. The results chosen tend to be semantically relevant to the query term, and this semantic relevance is what can be exploited, as discussed in the next subsection.

In this way, objects are grouped according to user agreement on their mutual relevance to a query term. Unlike that which happens in many semantic web and information management uses, objects are not grouped or classified according to whether they satisfy the predetermined membership constraints for a predefined structure of concepts. Using co-active intelligence, the actual concepts and their structure and interrelationships are themselves fluid and changing, in exactly the same way that language itself is. In short, *co-active intelligence systems favour*

observation over design.

Outline of co-active intelligence algorithms

Observation

The first stage is unobtrusive observation. We observe the behaviour of users as they interact with a normal web search engine. As they enter queries and scrutinise pages of results, we record their selections without impeding their normal searching behaviour. The click-through information we collect during this phase is

- *Anonymous*. Nothing ties a user to the data they provide.
- *Simple in form*. We need no more than a session identifier, the query term that was searched for and the web pages that were selected.
- *Easy to collect*. The process requires nothing more complicated than a proxy server placed between the user and the underlying search engine

There is no set duration for this observation phase. However, there is a definite correlation between the accuracy of the final results and the effort expended upon initial data capture, as discussed in section 5.1.

Sorting

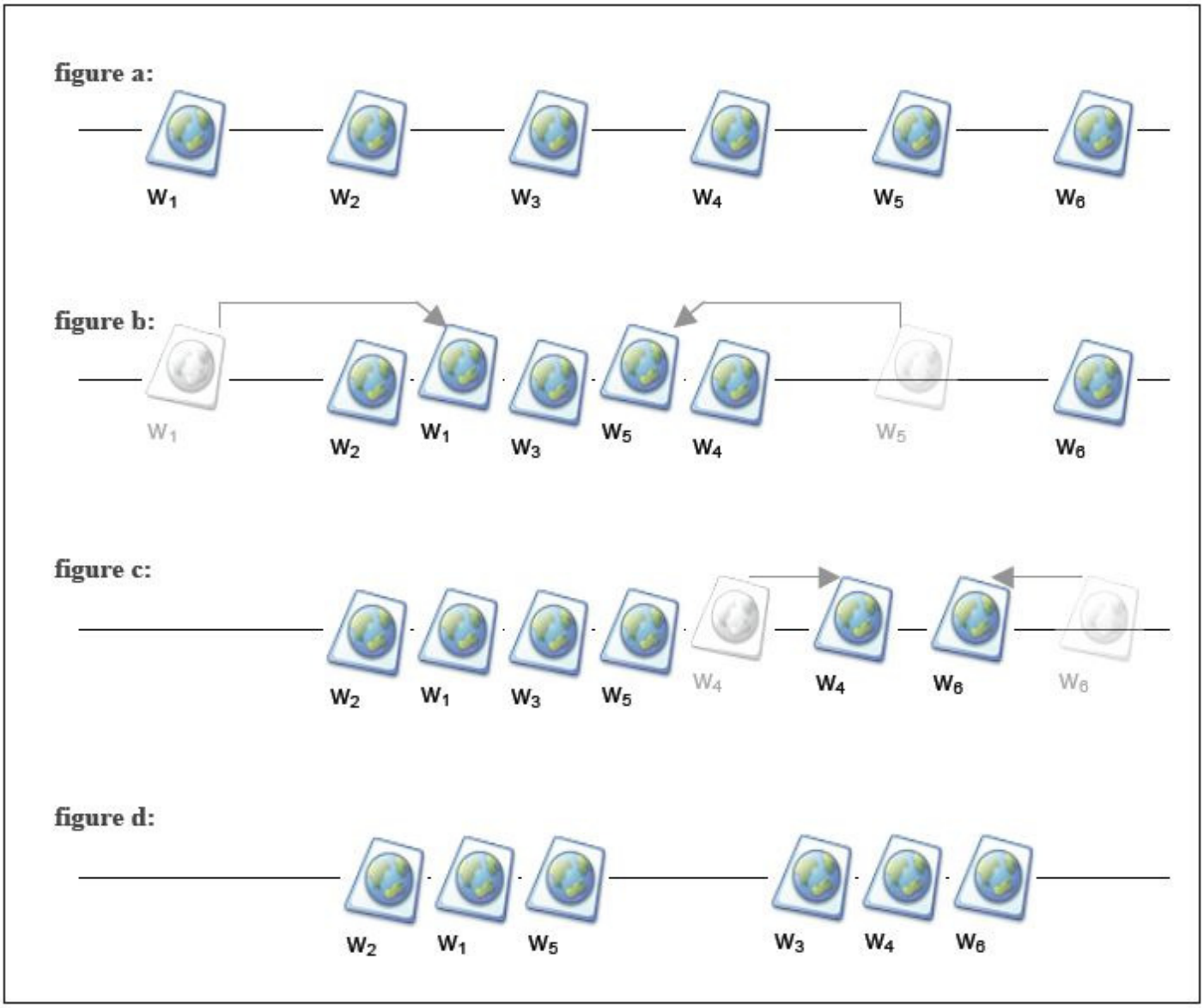
The next phase is a sorting process, which attempts to divide documents using a term in a particular sense from documents using the same term *in a different sense*. While this exercise is nothing short of routine for humans - who can quickly resolve different senses of the same word by drawing on context - sense discrimination has proven extremely difficult to automate (Krovetz & Croft 1989, Krovetz 1997). That is, until now!

To illustrate the sorting technique, assume the following conditions:

- Query term *x* submitted to a search engine returns a results list of six web pages *{w_1, w_2, w_3 ... w_6}*.
- Query term *x* is a homonym and has two different meanings *(x_1 and x_2)*
- Half of the web pages *{w_1, w_2, w_5}* that are returned when query term *x* is submitted concern *x1*
- Half of the web pages *{w_3, w_4, w_6}* that are returned when query term *x* is submitted concern x_2
- Some users are interested in x_1 and some are interested in x_2.

The next step, although simple in nature, is best illustrated visually.

1. We model the six web pages as points equally distributed along a straight line (see FIGURE a). Think of this line as a *spectrum* of the possible meanings *x* can have.
2. Now assume that a user of the search engine enters term *x* as a query term and then selects web pages w_1 and w_5 from the result list. Having recorded this activity, we move the points representing w_1 and w_5 closer together (see FIGURE b). Here the user's selections are being used as *implicit evidence* that these documents relate to the term in the same sense: See (Claypool *et al.* 2001, White *et al.* 2002).
3. Now assume that a second user enters term *x* as a query term, selecting web pages w_4 and w_6 from the result list. We record this activity and move w_4 and w_6 closer together (see FIGURE c)[2].
4. As more users search for term *x* and make their selections, this simple associative sorting algorithm converges the web pages into two sense-similar groups, each addressing a single sense (see FIGURE d).



Figures a-d: Stepwise illustration of an iterative sorting algorithm based on user behaviour.

Serving

In the final stage of the procedure, we combine the normal ranking information supplied by a traditional search engine with the information contained in these sense clusters. This allows us to present the user with coherent groups of web pages rather than the normal intermingled list. Overall, the technique reduces the cognitive load of processing the results, allowing a user to 'skim' the information rather than read it, pre-selecting document sets based on their overall meaning, rather than looking at all individual documents.

Any Media

The example above centres upon sense discrimination in relation to web pages, but co-active search techniques can be applied to search environments other than documents with little or no modification (for example, see (Truran *et al.* 2005) where the principle has been applied to image collections, as discussed in section 3.1 below). Wherever a search system provides *choices* to the user, and moreover has some mechanism for recording the *selections* from those choices tendered by the users, this technique can be applied. This means it can be used to partition audio files, images, video clips, spoken word fiction, application files etc.

RELATED WORK

This section covers three major areas of related work, the first looking at other co-active intelligence work, including word sense disambiguation applications, the second covering semantic web technology and the last being related work on the development of dictionaries.

Related work on coactive intelligence

So far, few efforts have been made to exploit the judgement of relevance implicit in user interactions. Interestingly, one of the earliest such efforts was in a commercial system, the Amazon online bookseller [HREF7], which provides a recommendation system based on its logs of purchases. Such recommendations are made at various stages, such as when the buyer logs in, browses a specific item, adds an item to their basket for purchase, or passes through the checkout. At each stage, they are offered additional items which may be of interest to them, based not only on their own past purchasing habits, but based also on the purchasing habits of numerous other customers. For example, if a number of buyers purchased item A as well as item B, then it is plausible to assume that items A and B have something in common, so that a new buyer of either of them would be likely to be interested in the other. While details on the success of this principle are commercially sensitive, it clearly functions well, as evidenced by the expansion of the service.

Interestingly, this recommender service knows nothing about books, and certainly nothing about their content. This might leave us wondering why it actually works. However, it works because it is just a cog in a bigger system, a hybrid of a processing algorithm and the situated intelligence of humans that interact with it. Its intelligence is purely an illusion, a mirror of all those people who interface with it - in other words, co-active intelligence. It is a form of artificial intelligence which gives apparently intelligent feedback, not operating in isolation like some AI systems, but as part of a wider interpretation of what an "intelligent system" can be.

Co-active intelligence is also the underlying principle of document ranking systems (Kleinberg 1999, Brin & Page 1998). The purpose of such algorithms is to detect "hubs" and "authorities", where hubs are pages that offer pointers to many other relevant documents and "authorities" are pages that many other sites judge to be, as the name suggests, authoritative on a given topic. Hubs are identified by the number of outgoing links while authorities are detected by the number of incoming links. Authorities in particular show how user consensus provides low-cost community judgement on the relevance and reliability of key documents or sites.

Co-active intelligence has more recently been turned to the quite difficult problem of image classification. Content analysis algorithms have some success but no software yet has the general ability to recognise and label images with the same accuracy as humans. For this reason, co-active intelligence is a potentially powerful tool for image classification, as it relies on the content-analysis abilities of groups of humans. If it can be implemented in a way that does not appear to impose unwanted tasks on the user, then this content analysis and image classification can be enabled at very little apparent cost.

Two quite distinct methods for doing just this are evident in the literature, these being the co-active intelligence search tool, SENSAI, developed by the authors (Truran *et al.* 2005), and a game developed to provide amusement to users to entice them into providing their judgement (von Ahn & Dabbish 2004). However what both approaches have in common is that they offer some related service that provides value to the user while collecting data on the user's selections non-intrusively.

The service offered by the SENSAI system is essentially a search engine that observes user selections from normal search engine and treats these selections as a form of relevance feedback on the results of the query. Normal search engines are unable to do this - they provide the set of results but are not usually part of the result-choosing process, as the user's browser goes directly to the selected results. SENSAI intercepts user choices transparently, and can initially run passively, merely collecting data on user queries and selections, but after collecting sufficient data, can begin to operate more actively, by ranking search results according to the relevance of those same results against the same or similar queries, as judged by previous users.

The service offered by the ESP game introduced in (von Ahn & Dabbish 2004) is an entertainment service, in this case offering the users a fun computer game. It also uses the principle of hidden activity recording. The image classification is more overt here, as the game requires users to guess what label their game partner has applied to an image that they both see. Thus labelling is a specific activity of the game and the data recorded by the system seeks out the most frequently used labels for each image.

The ESP game procedure is quite distinct from the Perpetual Dictionary's approach. The Perpetual Dictionary observes users entering in a word or phrase and selecting resources that are relevant to that word or phrase, and from this either directly or indirectly eliciting the meaning of the word or phrase. In contrast, the ESP game first provides the resources and asks the user to select a word or phrase that corresponds. In short, the Perpetual Dictionary is a directory of resources while the ESP game is a labelling service.

Probably the most relevant work to the Perpetual Dictionary is the Open Mind Word Expert (Chklovski & Mihalcea 2002)[HREF8]. Initially, the Open Mind initiative operated by requiring volunteer activity so did not operate behind a separate service but was populated explicitly and voluntarily (Stork 1999). However the Open Mind Word Expert now aims to solve the lexical ambiguity problem with user consensus by using the game approach typified in the ESP game (see above). Game participants are asked to discriminate word senses in a game, and their collective outputs select the meaning for words and phrases. However, there still remains a major distinction between the Word Expert and the Perpetual Dictionary - the Word Expert is forming a corpus of meaning that is similar to the operation of a dictionary, but it *clarifies* meaning for words or phrases *in a specific context*, based upon the extant set of meanings in an already-published dictionary. The Perpetual Dictionary does not provide static, fixed meanings, ever, but rather a pointer to a collection of resources whose membership varies, and from which users generate their own sense of the meaning for the word or phrase. In short, the Word Expert is a sense discriminator, not a dictionary.

Similar to the early Word Expert is the Wikipedia "Wiktionary"[HREF9], requiring explicit user contributions to populate the dictionary. While meanings are contributed by users, it relies rather less on consensus for its final results, since users are aware of and can dispute or alter each other's contributions. In some ways, it is closer to a formal dictionary, since fixed definitions are explicitly specified. Incorporating supporting documents, images and other media must also be done explicitly.

In summary, co-active intelligence exploits an increasingly important and accessible form of relevance feedback, allowing the emergent characteristics of masses of user interaction to be exploited in tasks that are either too labour-intensive for humans or currently too difficult for algorithms.

The Semantic web

In section 2.1, we noted that in co-active intelligence search, "objects are grouped according to user agreement on their mutual relevance to a query term", but in contrast that "in many semantic web and information management uses, objects are ... grouped or classified according to whether they satisfy the predetermined membership constraints for a predefined structure of concepts".

This is a key distinction between the semantic web and co-active intelligence. They share a common goal, of creating an infrastructure that allows mutually-comprehensible understanding of Web resources for retrieval and indexing purposes. However, the approaches are radically different.

The Semantic Web is defined to be "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee *et al.* 2001)[HREF10]. Note that the information is given something additional, namely metadata that describes the information. Note also that the additional material is well-defined meaning, so that it could be argued that the semantic web initiative aims to provide meaning for every addressable Web object eventually, and thus could be made to function as a dictionary in this sense.

The mutual comprehension is enabled through the Resource Description Framework (RDF), which provides a mechanism for specifying an agreed structure for a type of data, allowing it to be exchanged. Ontologies are instances of those agreed structures for sets or types of data for use in specific applications or areas, so could be seen as a form of glossary, acting as a dictionary in a limited domain.

This is where the greatest contrast with the Perpetual Dictionary approach can be seen. Developing an ontology is in principle similar to developing a relational database schema, and suffers from similar limitations. The most pressing limitation is that, having determined a particular structure/schema, and populated the metadata created for each object, the commitment to that structure becomes larger as the number of objects grows, and changing the structure inevitably encounters increasing resistance or inertia.

This is characteristic of the invariant structuring of relationships which occur frequently in data and knowledge management systems, not just semantic web and relational databases. The structure and membership constraints are defined ahead of time, by an authority (often a domain expert), and can be difficult to alter. For example, altering the schema of a relational database can be enormously difficult, because the data is logically stored within the structures. Likewise altering an ontology carries substantial overheads because the metadata in the format specific to

the ontology has been created. In fact, the resistance to change could be greater in semantic web applications because the mutual comprehensibility that is its aim has naturally engendered more applications relying on the agreed structure, whereas relational databases and their schema tend to be shared less often. Critics of the semantic web (Marshall & Shipman 2003) have suggested that this is one major reason why the semantic web is more suited to providing catalogues of products and services than to providing access to more abstract material.

Contrast this to the co-active intelligence principle of superimposing meaning on extant data in arbitrary structures. It might humorously be said that "no data were harmed in the making of this dictionary" because it is wholly unnecessary to alter, add to or in any way amend objects in order to assign them meaning. The meaning exists independently of any one object, and the binding between the meaning and objects is not committed, and in fact is expected and intended to change (see section 4). The objects are actually indicators for the meaning, and as the meaning changes with current usage, so too the participation of objects which indicate the meaning changes, with new objects arriving and old disappearing. There can be many meanings held in different locations and derived from different user populations, and all will be "correct" within their context.

In fact, the last decade has seen a trend toward the notion of superimposed structure, recognising that such superimposed structure, such as with the open hypertext systems, allow an unprecedented flexibility of structure, whereas inbuilt structure, as with relational database systems, favour efficiency but at the cost of flexibility. For example, it is now well-established in hypertext that the greatest flexibility arises from the separation of structure from data (Davis 1995), which not only allows non-invasive third-party data manipulation but permits, indeed expects, change in the underlying data by implementing a late binding of structure that can reflect the most up-to-date state of the data (Brailsford 1999). Co-active intelligence systems, like open hypertext systems, are designed to manage and exploit change.

A further issue that makes semantic web technology unsuitable for a global dictionary is that of bias, however unintended, on the part of the authority or domain expert whose responsibility it is to create the ontology. This is inevitable in any authority-based definition because the authority, while to some extent representative of the entire user group, usually cannot determine a structure that suits all the requirements of all who might wish to use it[3]. In a co-active intelligence system such as the Perpetual Dictionary, all voices participate, and while the greatest consensus creates the most visible results, lesser clusters of different meaning are still valid, and in fact can form secondary or localised/specialised meaning, and thus represent minority groups or communities.

Formal Dictionaires

In this section we look at the derivation of meaning through a co-active approach, as opposed through formal definition.

As discussed in the introduction, there may be differences in output from the Perpetual Dictionary and formal dictionaries, but there are also similarities in their purpose and use. Each aims to provide a service to a user who seeks a meaning or explanation of a term, and each aims to record and archive such meaning for a comprehensive part of a language. The two approaches may be different, but are in no way contradictory, and as noted below, can become merged.

Are formal dictionaries co-active?

A traditional dictionary enumerates static, explicit meanings for terms that are collected by some formal authority (such as the Oxford English Dictionary). Since these meanings are fixed at publication, traditional dictionaries begin to date as language evolves/regresses (see (Dent 2005)). This is exactly the same problem that causes some link integrity errors in a hypertext system (Ashman 2000) - publication is essentially a "snapshot" of the language at a specific time, and becomes increasingly incorrect as the snapshot ages and becomes unaligned with the current state.

Formal dictionaries do however incorporate change through the publication of new editions, as they are intended to archive meaning derived from usage, not by authority. It is true that formal dictionaries do themselves tend to become authorities, however language usage and dictionaries can be said to "co-evolve" because people develop new words or new usages[4], while still using the meanings already agreed previously and formalised in the dictionary, which itself is then updated to respond to those new words and usages. Thus a formal dictionary could certainly be argued to be produced through co-active processes, admittedly somewhat slow to react to

change in language, but nevertheless still based on evolving user consensus.

Distinctions in output

A co-active intelligence search engine fractionates a document collection to produce implicit, fluid meanings for terms that are tailored to the users of that search engine. So ultimately, co-active intelligence meanings are necessarily derived by the users. For example, a formal dictionary might define the term 'pancake' in the following terms:

> *A thin flat cake of batter, fried on both sides in a pan and usually served either flat (in the U.S. often with several stacked) accompanied by butter, syrup, lemon juice and sugar, etc., or rolled up with a sweet or savoury filling ...*
>
> *Aeronaut. A vertical descent made by an aircraft in a level position (e.g. as a result of a stall); spec. (more fully pancake descent, pancake landing) a landing in which an aircraft drops vertically after having levelled out close to the ground. ...*
>
> *A proprietary name for: foundation or other make-up in the form of a flat solid layer of compressed powder, widely used in the theatre.*[5]

In contrast, a co-active intelligence dictionary will produce 3 clusters of web pages, each corresponding to one of these senses. However, at no point is the term defined in black and white, unless a formal dictionary definition forms part of the results[6]. Instead, the user formulates their own understanding of the meanings of the term by looking at the documents in each cluster, in much the same way that normal dictionaries offer examples of a term's usage to help a reader clarify their understanding. As noted in section 1, a formal dictionary provides a definition with optionally some examples, whereas the Perpetual Dictionary provides examples which may include a formal definition from another dictionary.

THE PERPETUAL DICTIONARY

Determining meaning, or aboutness, is a natural application for co-active intelligence. Meaning itself is determined by mutuality, since communication breaks down when meaning is not shared between interlocutors [7]. If it is possible to focus on individual meaning by consensus, then an obvious extension of the principle is to create a complete, organic dictionary, whose content is determined by common usage. A dictionary based on co-active intelligence accurately reflects the way formal dictionaries are derived, with meaning determined by common usage, not by authority.

Formal dictionaries aim to provide a comprehensive listing of words and their meanings and usage, but inevitably language changes, and the publication of materials of any sort makes them susceptible to incompleteness when referring to changeable information. Substantial human effort must go into amendments and updates for creating new editions in order to reflect amended language. Finally there is the question of what goes into a dictionary- most dictionaries are limited by space and necessarily constrain themselves to a single purpose, such as meanings for a language, translations to another language, synonyms, glossaries or similar.

The Perpetual Dictionary is an obvious and patently useful application area for co-active intelligence. With user activity establishing concepts and meanings by consensus, it is possible to catalogue every single concept in any language, whether represented by word, image, symbol or any other means. Updates occur naturally and implicitly, as part of an ongoing process of assimilation of user activity, and the human effort required to establish meaning is distributed amongst the entire user population, and more importantly, without any additional effort required by them(In fact, not only is user effort not required, the users need not even know about the process for it to operate effectively [8]. And without space constraints, there is no reason that a single dictionary could not contain words in many languages, with many example uses and illustrative images. Images can also be catalogued in the dictionary and synonyms and translation between languages could be effected also (see section 5.2).

The Perpetual Dictionary would eventually have an entry for almost every concept, including all meanings. It can preserve older or archaic meanings, perhaps in a versioning system, so that a user can find out what something used to mean at any given time - for example, how many readers will remember gopher[9] and know what that word meant in the computer science context? Historians and students of language already know how language evolves, so that true understanding older documents requires awareness of those changes [10]. The same is also true of legislative documentation, as cases in law must refer to the law is it was at the time of the alleged offence. All known meanings and usages need to be preserved - something that is

increasingly difficult in a paper dictionary, and in one case, a complete dictionary of past and present words and technical terms in a single language runs to 20 volumes [11].

In short, anything used online can contribute to a multilingual, always up-to-date dictionary of what everything means, using co-active intelligence.

The meaning Emergent from coactive intelligence

Co-active search can determine one, any or all meanings of resources. The 'meaning' of a particular resource is inherently subjective, varying according to context of use, topic, author and reader bias and even date of use. Language, as well as non-verbal communication, is in a constant state of flux. As noted by Wittgenstein, "For a large class of cases - though not for all - in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language." (Wittgenstein 1953). So the meaning of a word or phrase is whatever its users think it means, in short the user consensus on that word or phrase.

Thus no system can with certainty determine the "true" meaning for a word, phrase, image or document, as there may not be any such thing. What co-active intelligence can do however, is to help discover one or more generally-agreed meanings.

What the system does is to mediate a meaning for a particular resource which is inferential in nature. We define an image by the images surrounding it in a particular space. We define a document by the documents arrayed about it along a particular dimension. The nature of 'meaning' in this context becomes somewhat convoluted. If we define a resource inferentially by examining its neighbors, and the meaning of the neighboring resources are defined in the same way, then we create a non-terminating self-referential circle.

However, a co-active search system encourages a patterning of a set of things along user-defined planes in which 'meaning' is

1. self-sustaining - it is generated with no outside help (other than continuing search-result selections of the users) and as a process it is self-constructing, self-organising and self-sustaining.
2. emergent - meaning is never determined by any one entity, but is implicit and arises from the collective activity of groups of users
3. never fixed - the Perpetual Dictionary is not ever meant to be a finished work, in fact, quite the opposite. Thus it is *perpetual*.

Essentially, the "feature space" of a co-active intelligence application, the area where relationships between items in the system are modelled (as exemplified in figures a-d of section 2.2.2), can be thought of as an implicit dictionary. It is not something that could be printed out, and the 'definitions' of each term are in fact pointers to resources about that term rather than literal descriptions, but the comparison holds true.

The Perpetual Dictionary, as an online resource, is implementable as a versioning system, where previous meanings can be preserved (as required with some versioning systems, especially legislative) so that it is possible to find out what meaning was current, and by implication, what meaning was intended, by the author at the time of writing. The version at any given time would essentially act as a "snapshot" of language at that time.

One particularly interesting feature of keeping "versions" of the Dictionary is the opportunity for historico-linguistic analysis. Snapshots of the feature space over time could provide evidence of the way in which the significance of words change over time. For example, the word 'wiki', which in 1996 was simply Hawaiian for 'quick', but now of course is heavily associated with a particular form of collaborative web authoring [12]. Query logs for the last decade could trace this linguistic drift. It would be a simple matter to represent and record language drift, as well as providing current meaning indicators for a given time.

IMPLEMENTING THE PERPETUAL DICTIONARY

In this section we overview the current state of the work and also consider other developments arising such as the possibility of exploiting the feature space to provide thesaurus and translational services.

Current Status

The SENSAI system discussed in section 3.1 above forms the core of the co-active intelligence that underlies the Perpetual Dictionary. Essentially the Dictionary is almost a by-product of the

original purpose of SENSAI, which was to group search results according to similar meaning. For a basic, up-to-date dictionary, it is not necessary to generate or accumulate any additional information, since a simple look-up of a term will return a cluster of documents whose aboutness corresponds to the term, according to the judgement of previous users.

SENSAI operates in two modes, namely *passive*, where it does not attempt to categorise users' search results, and *active*, where it does. The passive mode is necessary in order to generate enough user data to form statistically significant clusters during the active mode. While there is not at this stage enough data to draw conclusions about the effectiveness of co-active intelligence, we have performed a user trial which demonstrates that the algorithms are efficient enough to not penalise users during use.

A preliminary trial to measure the efficiency of the algorithms took place in 2005, and involved 43 volunteers. One of the aims of the trial was to establish whether the algorithms for grouping sets of search results automatically by sense were as efficient as just returning a mixed-sense set of the same results.

The testing suite for the trial was a simulated search environment in which volunteers carried out timed retrieval tasks. A control group of users worked through a set of retrieval tasks using a traditional results list interface, where documents relating to the various different senses of the query term were mixed together. A second group of users worked through an identical set of retrieval tasks, but were presented with a results interface in which documents were categorised according to the various senses of the query term (with SENSAI running in active mode). The time taken by each individual to complete the exercise was recorded, and average group times were then normalised and compared [13]. The difference between the normalised, average completion time for the control group and the group using SENSAI with sense categorisation enabled was found to be not significant. In other words, the sense categorisation algorithms did not impose an efficiency penalty.

However, this relatively small trial is inconclusive in terms of measuring the anticipated improvement of larger-scale co-active search systems, since while the co-active clustering algorithms operate, they are functioning over too small a dataset to offer conclusions about anything other than efficiency. It requires longer-term trials on larger datasets and user bases to determine what level of benefit arises from this form of sense categorisation, which intuitively ought to improve on other mechanisms for similar reasons that co-active recommender systems are so successful (see 3.1).

We are currently running ongoing trials that exploit Web log analysis to extract semantic associations, rather than recording user activity with a proxy. This gives us access to far larger quantities of data and early analyses have started to populate the resource clusters far more effectively, as well as exposing several interesting characteristics such as the dominance of Google as search engine of choice and the low average number of search terms.

Synonyms and translational services

We noted above that co-active intelligence could apply equally well to multiple media, including those which are often more challenging to automatically categorise otherwise (see 3.1). The same principle can apply to objects in other languages as well as in other representations.

Judicious use of co-active intelligence makes it possible to use detected meanings to find other words which have the same or a similar meaning. This could be a useful tool for effecting a thesaurus or translation service. Essentially the co-active intelligence solution treats all resources about a concept as equivalent, regardless of their origin and language, so that the set of resources returned includes relevant documents in other languages. This can arise either through similarity measures on documents and clusters, defined using established information retrieval algorithms, or via "bridging" activity between documents in different language, when multilingual users select documents from multiple languages from the result set.

Thesauri and translation services are technically very similar - the requirement is to find an alternative word or phrase with the same or very similar meaning. The only distinction is that thesauri do this within a language, translation across languages. There are however additional problems arising in cross-language retrieval, as discussed in the next section.

The problems of Cross-Language Information Retrieval

Cross-language information retrieval raises two key problems. These are

1. *the query translation ambiguity problem* - The query translation ambiguity problem occurs

when several translation words are provided for a source word by a particular translation tool, but only some of them are appropriate (Gao *et al.* 2002). For example, when translating 'pic' from French into English, bilingual dictionaries might suggest both 'crown' and 'pick'. Deciding which one of these terms is relevant in the circumstances seems easy for humans, but is complicated for machines.

2. *the out of vocabulary (OOV)* problem - this arises when a word or concept is present in one language but not in another. True translation is essentially impossible and the aim then is to find the word or phrase with the nearest semantic sense. This can be particularly challenging when the semantic sense of a phrase bears little relation to the meaning of its constituent words, as often happens with colloquialisms, for example "raining cats and dogs" which essentially means "raining heavily". Its derivation is not obvious, and its meaning can be confusing to non-native speakers. For non-co-active algorithms, phrases such as this require special exclusion or inclusion rules, whereas in co-active systems, if enough users recognise the true meaning, it will be correctly classified and false meanings excluded.

Solutions have been proposed for the query translation ambiguity problem, such as:

- Co-occurrence statistics, the hypothesis that correct translations should be co-occur in text while incorrect translations should tend not to co-occur. Given the possible target equivalents for two source terms can infer the most likely translations by looking at the pattern of co-occurrence for each possible pair of definitions. That is to say, closer words tend to have stronger relationships (Ballesteros & Croft 1998);
- Mutual information is a technology for selecting the best translations of a set of given query terms contain the translations that are mutually related or statistically similar with one another by a bilingual dictionary (Mirna 2000, Jang *et al.* 1999);
- Modelling using the N-best query translations. Given the initial query, the target documents are ranked using two different models, query-translation model, which generates the most probable term-by-term translations of a query and query-document model, which evaluates the similarity of documents and translations. The performance is measured through the set of N most probable translations of the query (Federico & Bertoldi 2002).

However the drawback of all these solutions is that they rely on external (and static) knowledge sources such as dictionaries or parallel corpora, which, as noted above, is problematic as language changes.

Approaches using co-active intelligence

There are at least two possible co-active intelligence based approaches for discovering synonyms and translations:

- One approach is to find clusters in the feature space that have a large overlap. A substantial overlap suggests that the query terms generating the clusters are semantically very similar, and we would aim to determine the probability that such overlapping clusters are synonyms via comparison with an established thesaurus for a significant quantity of terms.
- Secondly we can exploit the multilingual capability of the user community. An approach, effective mainly for translation, requires finding documents in other languages resulting from the same query term. This will necessarily involve a form of "bridging" where users with an understanding of similar terms in different languages select from query result sets that cover multiple languages. In fact, it is likely that some form of bridging is already present in order for the documents to be returned against the same query term, which could arise either through the explicit provision of alternatives by a human (for example, when new or foreign terms, or proper nouns are used in a document in one language, they are sometimes accompanied by a translation in another language (Zhang & Vines 2004)), or automatically where, say, an image has been labelled by many people using numerous languages. Since co-active intelligence has already been used to find aboutness of images (Truran *et al.* 2005), this approach has already been shown to be feasible. The words or labels used in different languages in this way can be represented as several different dimensions, forming a cross-language dimension. When the documents in that dimension exceed some threshold, it can used to represent the OOV term documents as well as the OOV translation itself.

Obviously it is possible to combine these with other, information retrieval-based similarity measures to give a stronger result.

In each case the work involves two-way comparisons, namely to discover coincident clusters then check their thesaurus-based similarity, but also the converse to compare clusters on terms known to be synonyms according to a thesaurus.

Depending on the existence of a true translation, co-active intelligence may not be able to translate exactly, but it can certainly provide a "sense" for the intended meaning from the set of documents retrieved. This may facilitate the movement of words between languages where no translation equivalent exists (as has already happened in English repeatedly).

CONCLUSIONS

There are many branches of research that seek to bypass or substitute human judgement with algorithms, and while many of these algorithms contribute toward the automatic classification and analysis of information, they can not in the foreseeable future replace human judgement, being unable to deliver the same accuracy and human-relevant perception. Until humans understand their own thinking processes well enough to replicate them in software, most such algorithms are necessarily going to be inadequate in some respect and their aim should not be to *supplant* but to *supplement* human activity.

Co-active intelligence is a principle that leverages the superior analytical and classification skills of humans. It can be applied in many ways, such as to recommender systems, document and image classification and search engines. In this paper we have discussed a new application whose eventual purpose is nothing less than the complete and ongoing classification and meaning provision of all Web artifacts, including non-Web artifacts that are described or duplicated on the Web, to create a Perpetual Dictionary. This is done using the classification capability of co-active intelligence to collate a selection of the most relevant documents to a word or phrase, given as a query term, where that collection of documents represents the meaning of the word or phrase.

So human intellect and judgement are at the beginning and end of the process - collectively, it forms the classification required to create the most appropriate and relevant result set, then individually it forms the perception of meaning from the gathered result set, with each individual deriving their own perception with the resources provided. The result set is likely to include formal meanings taken from one or more authoritative sources such as dictionaries and glossaries.

The effectiveness of co-active intelligence is already established, both commercially such as in recommender systems and for classification work (see 3.1). Its use for the creation of a Perpetual Dictionary and associated thesaurus and translational services is an obvious extension of this.

We give the final word to a lexicographer who recognises that authoritative dictionaries do not reflect the entire common usage of a language:

> *Being in the dictionary is not a badge of honor. People aren't limited to words I've managed to capture and pin down. A dog doesn't have to be registered with the American Kennel Association to be a dog. It still fetches your slippers; it just isn't pedigreed.*
>
> *-Erin McKean, American lexicographer, Denver Post, December 29, 2005 >*

REFERENCES

Ashman, H. L. (2000). "Electronic Document Addressing - Dealing with Change", *Computing Surveys*, 32 (3), p. 201-212.

Bailey, P., Craswell, N. and Hawking, D. (2000). "Dark Matter on the Web", In *Proceedings of the 9th International World Wide Web Conference*, poster track.

Ballesteros, L. and Croft, W. B. (1998). "Resolving Ambiguity for Cross-Language Retrieval", In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p.64-71.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001). "The Semantic Web", *Scientific American*, May.

Brailsford, D. (1999). "Separable Hyperstructure and Delayed Link Binding". *ACM Computing Surveys*. 31(4es). December.

Brin, S, and Page, L. (1998). "The anatomy of a large-scale hypertextual Web search engine", In *Proceedings of the Seventh International World Wide Web conference*, p.107-117.

Chklovski, T. and Mihalcea, R. (2002). "Open MindWord Expert: Creating Large Data Collections with Web Users' Help", *DLib magazine*, 8 (6), June.

Claypool, M. P., Wased, M., and Brown, D. (2001). "Implicit interest indicators", In *Proceedings of*

the 6th international Conference on intelligent User interfaces, p.33-40.

Davis, H. (1995). "To Embed or not to Embed", *Communications of the ACM*, 38 (8), p.108-109.

Dent, S. (2005). "Fanboys and overdogs: the language report", *Open University Press*.

Federico, M. and Bertoldi, N. (2002). "Statistical cross-language information retrieval using N-Best query translations", In *Proceedings of the 25th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, p.167-174.

Gao, J., Zhou, M., Nie, J., He, H. and Chen, W. (2002). "Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations", In *Proceedings of the 25th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, p.183-190.

Jang, M. G., Myaeng, S. H. and Park, S. Y. (1999). "Using mutual information to resolve query translation ambiguities and query term weighting", In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p.223–229.

Kleinberg, J. (1999). "Hubs, authorities and communities". *ACM Computing Surveys*, v.31 (4es).

Krovetz, R. and Croft, W. (1989). "Word Sense Disambiguation using Machine-readable Dictionaries", In *Proceedings of the 12th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, p.127–136.

Krovetz, R. (1997). "Homonymy and Polysemy in Information Retrieval", In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, p.72–79.

Marshall, C. and Shipman, F. (2003). "Which semantic web?", In *Proceeding of the 14th ACM Conference on Hypertext and Hypermedia*, p.57-66.

Mirna, A. (2000). "Using statistical term similarity for sense disambiguation in cross-language information retrieval", *Information Retrieval*, 2(1), p.67-68.

Stork, D. G. (1999). "The Open Mind Initiative", *IEEE Intelligent Systems and Their Applications*, 14(3), p.19-20.

Truran, M., Goulding, J. and Ashman, H. (2005). "Co-active Intelligence for Image Retrieval", In *Proceedings of the 13th annual ACM international conference on Multimedia*, p.547-550.

von Ahn, L. and Dabbish, L. (2005). "Labeling images with a computer game", In *Proceedings of the SIGCHI conference on Human factors in computing systems*, p.319–326.

White, R., Ruthven, I., and Jose, J. M. (2002). "The Use of Implicit Evidence for Relevance Feedback in Web Retrieval", In *Proceedings of 24th BCS-IRSG European Colloquium on IR Research. Lecture notes in Computer Science 2291*, p.93-109.

Wittgenstein, L. (1953) *Philospohical Investigations*, trans. G.E.M. Anscombe.

Zhang, Y. and Vines, P. (2004). "Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval", In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, p.162-169.

HYPERTEXT REFERENCES

HREF1
http://www.cis.unisa.edu.au/
HREF2
http://www.unisa.edu.au/
HREF3
http://www.nottingham.ac.uk/cs/
HREF4
http://www.nottingham.ac.uk/
HREF5
http://www.tees.ac.uk/schools/SCM/
HREF6
http://www.tees.ac.uk/
HREF7
http://www.amazon.co.uk

HREF8
http://www.teach-computers.org/word-expert/english/
HREF9
http://www.wiktionary.org/
HREF10
http://www.w3.org/2001/sw

FOOTNOTES

1 But not always, as users can be distracted by interesting but irrelevant results.

2 The user's selections are implicit evidence that these documents relate to the term in the same sense: See (Claypool *et al.* 2001, White *et al.* 2002).

3 Colloquially it might be said: "you can't please all the people all the time".

4 See for example the additions to the Oxford Dictionary at http://www.askoxford.com/pressroom/archive/coed11new/?view=uk

5 Definitions taken from the Oxford English Dictionary (Second Edition), CD-ROM version 3.1, Oxford University Press, Northants., United Kingdom, 2006.

6 This is of course increasingly likely to happen where such definitions exist, as the coverage of the Perpetual Dictionary grows.

7 Just see for example, Alice's conversation with Humpty Dumpty in *Through the Looking Glass* by Lewis Carroll:

> *"When I use a word," Humpty Dumpty said in a rather scornful tone, "it means just what I choose it to mean --- neither more nor less."*
> *"The question is", said Alice, "whether you CAN make words mean so many different things."*
> *"The question is," said Humpty Dumpty, "which is to be master --- that's all."* .

8 In fact, not only is user effort not required, the users need not even know about the process for it to operate effectively.

9 See http://en.wikipedia.org/wiki/Gopher_protocol if you are too young to remember!

10 A classic example is the dialogue between Hamlet and Ophelia in Shakespeare's Hamlet Prince of Denmark, (Act III Scene 2) where a sequence of coarse and sexually-explicit remarks from Hamlet generally go unnoticed by modern audiences.

11 The Oxford English Dictionary, published by the Oxford University Press.

12 The word *wiki* is also still absent from many dictionaries, including the MS Word 2004 for Mac version 11.0.

13 To avoid confounding factors such as IT familiarity or previously acquired searching skill influencing our results, the retrieval task timings were normalised against a set of data collected prior to the trial in which the groups completed identical exercises.

COPYRIGHT