Portable Extraction of Partially Structured Facts from the Web

XXXXXXX
xxxxxx

Abstract. A novel fact extraction task is defined to fill a gap between current information retrieval and information extraction technologies. It is shown that it is possible to extract useful partially structured facts about different kinds of entities in a broad domain, i.e. all kinds of places depicted in tourist images. Importantly the approach does not rely on existing linguistic resources (gazetteers, taggers, parsers, etc.) and it ported easily and cheaply between two very different languages (English and Latvian). Previous fact extraction from the web has focused on the extraction of structured data, e.g. (Building-LocatedIn-Town). In contrast we extract richer and more interesting facts, such as a fact explaining why a building was built. Enough structure is maintained to facilitate subsequent processing of the information. For example, this partial structure enables straightforward template-based text generation. We report positive results for the correctness and interest of English and Latvian facts and for the utility of the extracted facts in enhancing image captions.

Keywords: Fact extraction, multilingual, information retrieval, information extraction, web, image captioning.

1 Introduction

This paper proposes a novel fact extraction task which fills an important gap between current information retrieval (IR) and information extraction (IE) technologies in order to further exploit the vast quantities of multilingual information available on the web. Search engines retrieve relevant web pages across diverse domains and across languages, but the onus is on the user to read through and interpret the results. By contrast, IE systems provide structured facts and data from natural language texts which are amenable to further automated analysis, and multi-document summarization systems and question answering systems fuse information about an entity or topic of interest to cut down reading time. However, such systems are typically costly to port to new languages and the domains in which they work tend to be narrow and comprise only a small set of entity types and relations. We believe that there are emerging applications, such as automated image captioning and augmented reality, which would benefit from exploiting information on the web across broad domains and multiple languages, but which do not require fully structured information or the majority of all available information about an entity. For example, to automatically enhance an image caption we only require one interesting fact about the place in the image, with enough structure for the fact to be inserted appropriately into a text generation template. In sacrificing the requirements for full structure and comprehensive information about an entity, we expect to gain considerably in coverage of domains and ease of porting between languages.

We elaborate these points in Section 2 as we define the ‘Tell Me About...’ task which is, in broad terms, to provide one or more of the most interesting facts about a given entity in a partially structured form that enables some further processing and re-use of the information. Section 3 discusses related work in the fields of IR and IE, with a focus on information extraction from the web, multi-document summarization and question answering. Section 4 presents a highly portable solution for extracting partially structured facts that exploits information redundancy on the web, i.e. the fact that the same information about an entity is available in many forms on the web. The crucial assumption is that at least one key fact about an entity will be expressed somewhere on the web in a simple form, so we only need to work with a few simple linguistic structures and very shallow language processing. We report positive results for the correctness and interest of English and Latvian facts (128 facts each judged by an investigator and five subjects). The utility of the “Tell Me About...” task is demonstrated by enhancing the captions of tourist photographs using extracted facts for template-based text generation, with an evaluation of caption readability (90 image captions each judged by six subjects). In closing, Section 5 considers the potential for generalising the task and solution to other domains and applications, and raises associated research questions.

2 The “Tell Me About…” Task

Let us elaborate on the details of this task, and the motivation for it, by considering one potential application – automatic image captioning. The number of digital images being archived in personal collections and shared in social image collections (such as www.flickr.com and www.panoramio.com) is increasing very rapidly. When users view images from these collections it is desirable to have information describing each image available in a caption. However, people taking pictures will often either not know sufficient details about the place depicted in the image to do this effectively or will not take the time to do this, so automated solutions are required. There is also a burgeoning interest in augmented reality whereby a camera screen on a mobile device is updated automatically with caption-like information about the place that the camera is pointed at. Digital image capture devices are increasingly incorporating location sensing via GPS monitoring. This can be combined with other image metadata such as the date and time of capture and cross-referenced with geographic databases to generate simple descriptive captions for an image, e.g. of the form “North Bridge photographed in the afternoon” [1]. We see an opportunity to exploit the vast information content of the web in order to enhance such a caption with a key fact, e.g. to output something like “North Bridge, which was built to link the New Town with the Old Town, photographed in the afternoon”.

Whilst we can be confident that information about many places is available in many languages on the web, the challenge is to identify the most interesting key facts for a given entity. There is also the challenge of extracting information into partially structured facts that enable further processing and re-use of the information. In the image captioning scenario simply adding whole sentences from the web to an existing caption would have unpredictable results for caption readability. It could be that a long sentence contains information about more than one place, so we need to identify just the relevant part of the sentence. Also, if we want to insert information into an existing caption, i.e. into the middle of a sentence, then we need to know something about how it phrased. For the "Tell Me About..." task we specify that facts should have the form of a triple – (Entity, Cue, Text-Fragment), where 'Cue' is one of a fixed set of information cues (loosely akin to relations), and 'Text-Fragment' is a text fragment taken directly from a webpage, such that 'Cue Entity Text-Fragment' reads naturally as a sentence, e.g. (North Bridge, was built, to link the New Town with the Old Town). For template-based image captioning this means we can, for example, insert information in a subclause starting with "which" for cues such as "was built", but removing "which" and the cue itself for cues like "is". The partial structure of the fact gives us control over text generation that we would not have if the fact was only a text fragment. However, because the right-hand side of the fact is a text fragment, and not another entity of fixed type (as it would be in a standard IE template), then the same cue can get quite different kinds of information, allowing for much richer facts when available, e.g. (Hadrian's Wall, was built, in AD 122-130 on the orders of the Emperor Hadrian), (Hadrian's Wall, was built, to keep out the marauding Scottish).

To summarise, the "Tell Me About…" task proposed here is as follows. Given the name of an entity, and a specified language, a list of facts about the entity should be returned in the form (Entity, Cue, Text-Fragment) and sorted so that more interesting facts come higher; the precise notion of 'interesting' may vary from application to application. With regards to the image captioning scenario, it is important to note that the place depicted in a photo may be one of very many different kinds of entity (bridge, monument, beach, church, mountain, statue, glacier, plaza, etc.). Furthermore, the most interesting aspect of one entity may not be the same as the most interesting aspect of another entity of the same type – one church has spectacular stained-glass windows, another is known for an historical event that happened there, a third offers amazing views from its tower. Finally, a caption for an image on a website may be required in many languages. For these reasons, as we discuss next, current IE approaches are not appropriate.

3 Related Work

Although we consider "Tell Me About…" to be distinct from other natural language processing tasks, it does clearly have similarity with established and well understood tasks within IR and IE. The idea of ranking facts could be seen as similar to the ranking of documents for IR [2], and, more specifically, the retrieval and ranking of passages [3]. Indeed, snippets returned by web search engines are the starting point in our approach to fact extraction, although by the end of the process the sorted facts are in a different order than the snippets ranked by the search engine. The extraction of 'partially-structured' information makes our fact extraction look quite a lot like information extraction [4], but whilst we do specify a set of cues (similar to relations), we do not require the structuring of the right-hand side text fragment into a template (which could, for example, make relations between entities explicit). We have found that this makes it possible to pursue quite a generic approach to fact extraction across broad domains and multiple languages, whereas IE systems require non-trivial amounts of work to be adapted to different kinds of entities and languages. Question answering systems return facts, typically in response to factoid questions with answers that are dates, locations, organizations, people, etc. [5]. However, for a given entity it is not possible to anticipate what, if any, factoid question will give the most interesting information. That said, our approach to fact extraction shares some features with shallow question answering methods that exploit redundancy in the snippets returned by web search engine, e.g. [6]. This leads to two key assumptions: (i) the same information is expressed in many ways across the web, so it is only necessary to look for it in a small number of relatively simple forms; and, (ii) overlaps between what is written on different web pages can be used to compute an 'interest' score for facts. Multi-document summarization systems do something rather like the "Tell Me About…" task when they select a set of informative sentences about an entity, e.g. [7], but with a focus on more than just a few key facts, and the need to produce coherent text as output, such systems typically depend on quite extensive linguistic resources – at a minimum training corpora – that mitigate against porting easily between languages.

Previous work on information extraction from the web, rather than from domain-specific collections of a single text type, has achieved impressive quantities of facts at high levels of precision, e.g. 1 million ranked facts with a pre-specified relation at 75-98% Precision [8]. Under the rubric of 'open information extraction', which discovers relations as well as facts, a precision of 88% has been reported [9]. In related work the TextRunner system extracted over 500 million tuples from 120 million web pages [10]. However, much of this previous work has focused on the extraction of wholly structured data to specify relations between two entities, e.g. facts of the form (City-CapitalOf-Country), (Person-BornIn-Year), or (Company-Acquired-Company). Whilst this effectively enables the storage, analysis and retrieval of millions of facts in relational databases, these relatively simple facts are unlikely to be interesting for applications such as image captioning. An online demonstration does suggest that the TextRunner system [11] can provide facts with unstructured right-hand sides but our impression is that low quality of information is the price for exceptionally broad coverage. Furthermore, with regards to portability between languages, the approaches described by [9] and [10] rely on a linguistic analysis of how relations are expressed in English, and on syntactic parsers. Although the approach in [8] avoids syntactic analysis and parsing, it nevertheless works with text that has been part-of-speech tagged and draws on existing word distribution data. Taggers, parsers, and other commonly used linguistics resources are not available for most of the world's languages and so we are interested in an approach that does not rely on such things.

4 Our Approach to Fact Extraction

Here we present a first approach to the "Tell Me About…" task. We show how, given an entity (in this case any kind of place), we return a list of facts in the form (Entity, Cue, Text-Fragment), ranked according to a score which is intended to promote interesting and true facts. The approach is generic across a broad range of entities, and requires minimal effort to port between languages. Crucially, we assume that at least one key fact about an entity will be expressed somewhere on the web in a simple form, so we only need to work with a few simple linguistic structures and shallow language processing.

4.1 Algorithm for Fact Extraction

For a given entity, the following steps are followed to generate a list of facts about it.

I. Get Snippets from Search Engine. A series of queries is made to a web search engine (we used Yahoo's BOSS API [12]). Each query takes the form <"Entity Cue">; the use of double quotes indicates that only exact matches are wanted, i.e. text in which the given entity and cue are adjacent. A set of cues is manually specified to capture some common and simple ways in which information about the general kind of entity is expressed. For places we used cues like 'is a', 'is famous for', 'is popular with', 'was built'. Although we worked with around 40 cues (including single/plural and present/past forms) it seems that a much smaller number are responsible for returning the majority of high ranking facts; in particular (and perhaps unsurprisingly) the generic "is" seems most productive. The query may also include a disambiguating term. For example, streets and buildings with the same name may occur in different towns, so we can include a town name in the query outside the double quotes, e.g. <"West Street is popular with" Bridport>. For each query, all the unique snippets returned by the search engine (up to a specified maximum) are processed in the next step; typically a snippet is a few lines of text from a webpage around the words that match the query, often broken in mid-sentence.

II. Shallow Chunk Snippets to Make Candidate Facts. Because we are only retrieving information about our given entity that is expressed as "Entity Cue …", then we can use a simple extraction pattern to obtain candidate facts from the retrieved snippets. For both English and Latvian the gist of the pattern is '`BOUNDARY ENTITY CUE TEXT-FRAGMENT BOUNDARY`', such that '`TEXT-FRAGMENT`' captures the 'Text-Fragment' part of a fact. The details of the pattern are captured in a regular expression on a language-specific basis, e.g. to specify boundary words and punctuation, to allow optional words to appear in between `ENTITY` and `CUE`, and to reorder the elements for non-SVO languages. A successful match of the pattern on a snippet leads to the generation of a candidate fact: using the extraction pattern in the Appendix, the snippet text '...in London. Big Ben was named after Sir Benjamin Hall. ...' matches, giving the candidate fact (Big Ben, was named, after Sir Benjamin Hall) but 'The square next to Big Ben was named in 1848...' does not match.

III. Filter Candidate Facts. Four filters are used as a quality control, the first two of which require the specification of language-specific word lists which were built manually over a number of runs of the algorithm.

General filter words – a candidate fact containing any of the given filter words is removed; this can be used to remove potentially subjective statements containing ‘me’, ‘my, ‘our’, ‘amazing’, ‘fantastic’, etc.

Invalid end words – to catch some erroneous shallow chunking (most likely due to noisy web data, or to a badly cut search engine snippet) this filter removes candidate facts ending in words such ‘to’, ‘from’, by’, etc.

Length of Text-Fragment – a threshold can be set to filter out candidate facts with text-fragments shorter than the specified number of words; it seems that shorter text-fragments are more likely to lead to incomplete or incorrect facts.

Words all in capitals – when this filter is turned on, any candidate fact containing a word that is all in capitals is removed; this is good for removing spam and content in an informal style, but of course it also removes candidate facts containing acronyms.

IV. Score and Sort Facts. Our idea here is to rank facts, at least coarsely, so that we are more likely to get correct and interesting facts at the top. The notions of correctness and interest are each problematic and difficult to unpick for the purposes of algorithm design and evaluation. Here we exploit the overlap between candidate facts for the same Entity-Cue pair to capture these notions to some extent. For each Entity-Cue pair a keyword frequency list is generated by counting the occurrence of all words in the Text-Fragments for that pair; words in a stop word file are ignored. The score for each fact is then calculated by summing the Entity-Cue frequencies of each word in the Text-Fragment, so that facts containing words that were common in other facts with the same Entity-Cue will score highly. If shorter facts are wanted then the sum is divided by the word length of the Text-Fragment. We see two main ways in which the sum score for a fact can get high: (i) there are many overlapping Text-Fragments for an Entity-Cue pair, so there are some high word frequencies; and, (ii) a fact contains more of those high frequency words than other facts. Thus, we hope to get high ranked facts with the most appropriate Cue for the Entity, and the best Text-Fragment for the Entity-Cue pair.

To give an impression of how ranking works, Figure 1 shows the top and bottom 10 facts returned for ‘Eiffel Tower’, using the ‘sum only’ scoring. The top ranked facts are generally rich in correct information about the given entity. In contrast, incomplete and trivial facts end up low down the list. We see that 4 of the top 10 facts have the Cue “was built” which seems like a good cue for interesting information about an historical monument. The high-ranking facts with this Cue include words like “Paris”, “1889”, “international”, “exhibition” which are likely to appear after “Eiffel Tower was built…” on many web pages – the fact with all four of these words is ranked highest. For Latvian the top-ranked fact was “Francijas pazīstamākajiem simboliem un gadā to apmeklē aptuveni seši miljoni cilvēku”, which translates to “The Eiffel tower is one of best known symbols of France and it is visited by around 6 million people a year”; this suggests that there is less information about the tower’s history available on the Latvian web.

For an example of how an undesirable fact is ranked low, see the fact ranked tenth from bottom in Figure 1. This includes the words “Paris” and “exposition” which generally would be highly associated with “Eiffel Tower” but since the fact has the Cue “is built” (rather than “was built”) then these words and the fact score low.

```
(Eiffel Tower, was built, in 1889 for an international exhibition in
Paris)
(Eiffel Tower, was named, after an ingenious engineer whose design of
the tower turned it into a reality and pride of the French nation)
(Eiffel Tower, is, an iron tower built during 1887-1889 on the Champ de
Mars beside the Seine River in Paris)
(Eiffel Tower, was one of, the first tall structures in the world to
contain passenger elevators)
(Eiffel Tower, was one of, the landmarks visited by Luigi when he came
to save Paris from invading Koopa Troopas)
(Eiffel Tower, was built, by Gustave Eiffel for the International
Exhibition of Paris of 1889 commemorating the centenary of the French
Revolution)
(Eiffel Tower, was one of, the first structures in the world to have
passenger elevators)
(Eiffel Tower, was built, in 1889 for the Universal Exposition
celebrating the centenary of the French Revolution)
(Eiffel Tower, was built, as a temporary structure for an exhibition in
1889)
(Eiffel Tower, is named, after its designer and engineer Alexandre
Gustave Eiffel)
...
(Eiffel Tower, is built, for the Paris exposition)
(Eiffel Tower, was famous, enough for everyone to know)
(Eiffel Tower, is made, up of a base)
(Eiffel Tower, was made, for the Exposition Universelle)
(Eiffel Tower, is made, of over 10)
(Eiffel Tower, is made, from 18)
(Eiffel Tower, is made, of 3 platforms)
(Eiffel Tower, is made, with 2)
(Eiffel Tower, is famous, throughout the world)
(Eiffel Tower, is famous, for a reason)
```

Fig. 1. The top 10 and bottom 10 facts for the entity “Eiffel Tower”.

```
(Highgate Cemetery, was opened, in 1839 in response to a lack of burial
   spaces in London proper)
(Highgate Cemetery, was opened, in 1839)
(North Bridge, was, originally built in 1772 to connect the burgh with
   the Port of Leith to the north)
(North Bridge, was built, to link the New Town with the Old Town)
(Bahnhofstrasse, is, where well-heeled bankers and perfectly-coiffed
   ladies shop for designer clothing and gold watches)
(Bahnhofstrasse, is, Zürich's main shopping avenue)
(Durdle Door, is a, natural limestone arch on the Jurassic Coast near
   Lulworth in Dorset)
(Durdle Door, is a, limestone arch)
(The Matterhorn, was one of, the last Alpine mountains to be ascended
   due to its imposing shape and unpredictable weather)
(The Matterhorn, was, first climbed in 1865)
```

Fig. 2. Pairs of facts about places. The first is the top ranked fact with simple sum score, the second is top ranked with sum / number of words.

To look at how we can select between long and short facts, Figure 2 shows the top ranked facts for a variety of places using the two scoring options described in IV, i.e. 'sum of word frequencies', and 'sum / number of words'. It seems from these examples that we have good control over fact length. Also, it is interesting to note that, in each pair, the long and short facts (which come from different web pages) give a similar kind of information even if they do not share many or any words.

4.2 Evaluation of Fact Extraction

For this evaluation we selected 68 place names in English and 60 place names in Latvian from around Europe. We chose an even mixture of urban / rural and famous / not famous places from European cities (London, Riga, Zurich and Dublin) and countryside (UK, Latvia, Switzerland and Ireland), and various types of place – churches, statues, mountains, rivers, etc. For each place the top ranked fact was used for evaluation; see Appendix for the settings used to generate facts.

Evaluating Correctness of Facts. Each of the facts in English was rated as correct or incorrect by an investigator by searching for the fact on the web in the following manner. If the fact was found on Wikipedia, or an official tourist website for the region, and on one other website, or if the fact was found on three independent websites, it was marked as correct. If part of the fact was found on the web using this technique, then the fact was marked as partially correct. Otherwise the fact was marked as incorrect. Due to the lack of coverage in Latvian on the web, Latvian facts were rated as correct if they were located on Wikipedia, or an official tourist website for the region or if they were known to be correct by the investigator.

For the English experiment 35 of the 68 facts were marked as correct (51%), 13 were partially correct (19%) and 20 (30%) were incorrect. Analysing the partially correct facts revealed that 11 of the 13 were incomplete facts, e.g.: (*Dridzis Lake, is, the deepest lake not only in Latgale*); here it looks like our chunking pattern cut too soon (i.e. on the word 'but'), although a similar problem occurs occasionally with the way the search engine creates snippets. The other two partially correct facts had spurious material at the end of the fact, e.g.: (*Mount Titlis, is the largest, winter sports paradise in Central Switzerland _ even the most demanding skiers*); the unusual punctuation '_' is missed by our chunking pattern. Analysing the 20 incorrect facts, we found that only six of them were actually false, for example a fact which was supposed to be about the National Museum in Zurich actually referred to a museum in Prague; this is despite our use of Zurich as a disambiguating term. Eight of the incorrect facts were unreadable, for example: (*Daugava river, is, soon to be a prelude of things to come that would prove 2000 wasn't Cappellini's year*) which we put down to web 'noise'. The correctness of the remaining 6 incorrect facts was actually indeterminable, e.g.: (*Bastejkalns Park, was, renovated during last winter*); we have since added words with temporal reference like 'last' to the filter words list, as well as deictic words like 'this'. Similar to the English results, for the Latvian evaluation 32 of the 60 facts were marked as correct (53%), 19 (32%) were partially correct and 9 (15%) were incorrect.

Evaluating the Interest of Facts. Ten native English speakers were each presented with 34 English facts to rate. Ten native Latvian speakers were each presented with 30 Latvian facts to rate. In this way each fact was rated by 5 subjects. The lists of facts presented to subjects were randomly chosen using a Latin square. For each fact, subjects answered "yes" or "no" to the question: "Is this the type of fact you would expect to read in a travel guide?" The question is intended to get at the notion of 'interest' in a way specific to our application scenario, i.e. we assume users would be happy with travel guide like facts added to their image captions. Results are summarised in Table 1 which indicates that, more often than not, our algorithm is producing as its top ranked fact something that most people find acceptable as a fact for something like a travel guide.

Table 1. Responses from 5 subjects for 68 English facts, and 60 Latvian facts.

	English		**Latvian**	
	5/5 subjects said 'Yes'	**>= 3/5 said 'Yes'**	**5/5 subjects said 'Yes'**	**>= 3/5 said 'Yes'**
Is this the type of fact you would expect to read in a travel guide?	26/68 (38%)	53/68 (78%)	14/60 (23%)	38/60 (63%)

Our evaluation criteria for fact correctness were rather strict, which is highlighted by the fact that a majority of subjects rated more facts as 'interesting' (78% English and 63% Latvian) than we ourselves rated as correct (around 50% for both). As noted, it seems that some relatively simple changes to our extraction patterns and word lists will improve our 'correctness' score quite considerably, so overall we are confident that the fundamentals of the approach are sound. Importantly, the approach was very cheap to port between languages. In going from English to Latvian all that was required was a small modification to the extraction pattern, and the translation of the cue set (see Appendix); the other word lists were also translated, but for Latvian these did not have much impact on results.

4.3 Enhancement of Image Captions

Our initial motivation for doing fact extraction was to be able to add information about places into image captions. This provides an application scenario for evaluating the utility of the facts that we can extract. Lists of 30 image captions were created in English and Latvian for images depicting urban and rural places occurring in Ireland, UK, Latvia and Switzerland. Half the captions were in the form: "*Place* photographed in *location.*" The other half were in the form: "Photo taken near *place* in *location.*" Half of the captions also had the time of day inserted into the sentence, for example "Photo taken in the afternoon near *place* in *location*".

Each of the 30 English captions had a fact added in two different ways: 1) insert fact as sub clause in original sentence; and 2) append fact as new sentence to original caption; this led to 60 English enhanced captions. For (1) the string '`, which CUE TEXT-FRAGMENT`,' was inserted after the place name in the caption. For (2) a second sentence was formed by adding "`PLACE CUE TEXT-FRAGMENT`" after the original caption. Insertion as subclause was deemed inappropriate for Latvian, so we had just 30 Latvian enhanced captions with facts added as new sentences. In all cases we manually ensured that only correct facts were added because we wanted to concentrate on evaluating the readability of the enhanced captions.

The 60 English enhanced captions were presented to 6 native English speakers, in random orders for judgment. The 30 Latvian enhanced captions were presented to 6 native Latvian speakers. For each enhanced caption, subjects answered ‘yes’ or ‘no’ to the question: “Does this sentence read naturally to you?”. When facts were added as new sentences then a majority of subjects deemed 29/30 (97%) of the enhanced image captions to be readable both for English and for Latvian. The results (English only) for inserting facts as subclauses seemed to depend on the form of the original caption. For 15/15 (100%) of captions with the form “*Place* photographed in *location*” a majority of subjects judged the enhanced caption with fact inserted as subclause to be readable. Recall, we are only able to insert information as a subclause – which keeps the captions more compact – because we have partially structured facts (cf. Section 2). For the other caption form only 7/15 (47%) enhanced captions were judged readable by a majority of subjects; upon inspection, it seemed that these captions tended to be quite long already (including additional temporal information), so a further subclause, even though grammatically correct, became awkward to read.

5 Conclusions

To summarise, a new kind of fact extraction task was defined, and a solution to the task was evaluated for two very different languages. It was shown that it is possible to extract useful partially structured facts about different kinds of entity in a broad domain, using a common approach that ports very easily between languages in the absence of existing linguistic resources. In contrast with traditional IR techniques we produce output that is more amenable to further automated processing. In contrast with traditional IE techniques our approach has the potential for efficiently covering much broader domains and many more languages.

Of course we need to try other kinds of language before making strong claims about portability. Although Latvian is a free word order language, the SVO order does dominate, so we were able to get good results with just one extraction pattern. However, even in languages with more variation in word ordering, we expect that we could use just a few extraction patterns based around cue sets. What is less clear to us is the ease with which we can port to other domains. Whilst we found interesting facts about many different kinds of places were expressed using a relatively small number of common cues, this may not be the case for all kinds of entities. That said, in some very preliminary work, we were able to get some encouraging looking English facts about people and organizations using just a few cues.

Beyond the image captioning application and template-based text generation, we see potential for the “Tell Me About...” task in other areas. For some kinds of queries to search engines, users may benefit from being presented with a few facts about their topic of interest: we feel that our chunking of information and ranking of facts can add value to the snippets returned by a search engine. Recently some search websites have started to offer something more like ‘knowledge retrieval’ on top of information retrieval [13], [14], and our impression is that our kind of fact extraction could contribute to such endeavours. For example, given a number of entities of the same type, partially structured facts could be presented in a table to compare the entities.

Moving on, we want to develop a more thorough understanding of how, why and when the algorithm works, and to conduct more testing of the assumptions underlying the approach. It would be interesting to know more about how much of the available information our cue sets and simple extraction patterns retrieve and extract, and to know what cues are the most effective. Effective cues would be those that not only retrieve the most information from the web, but that retrieve information that is amenable to our shallow chunking and scoring techniques. We also plan to explore the potential for automatic cue generation by applying techniques from work done on pattern identification in unsupervised information extraction. With regards to the ranking of facts, more could be done to articulate and code the notions of ‘interest’ and ‘correctness’ although we are restricted to some extent by the difficulty in producing a gold standard evaluation set, and hence the need to present results from each evaluation run to subjects. Finally, we would like to explore the use of a tiling algorithm to merge similar facts (removing repetition) and to assemble larger facts from overlapping text fragments [6].

References

1. Purves, R. S., Edwardes, A. J & Sanderson, M.: Describing the Where – improving image annotation and search through geography. In: First Intl. Workshop on Metadata Mining for Image Understanding (2008)
2. Baeza-Yates, R. and Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, New York (1999)
3. Salton, G., Allan, J. and Buckley, C.: Approaches to passage retrieval in full text information systems. In: Procs. 16th ACM SIGIR, pp. 49--58 (1993)
4. Sarawagi, S.: Information Extraction. Foundations and Trends in Databases. 1(3), 261--377 (2008)
5. Lin, J.: An Exploration of the Principles Underlying Redundancy-based Factoid Question Answering. ACM Trans. Information Systems. 25 (2), 1--55 (2007)
6. Dumais, S. et al.: Web Question Answering: Is More Always Better? In: Procs. 25th ACM SIGIR, pp. 291--298 (2002)
7. Goldstein, J. et al.: Multi-document Summarization by Sentence Extraction. In: Procs. NAACL-ANLP 2000 Workshop on Automatic Summarization, pp. 40--48 (2000)
8. Pasca, Marius, et al.: Organizing and Searching the WorldWideWeb of Facts - Step One: the One-Million Fact Extraction Challenge. In: Procs. 21st Nat. Conf. on AI (AAAI-06), pp. 1400--1405 (2006)
9. Banko, M. and Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction. In: Procs. ACL-08, pp. 28-36 (2008)
10. Etzioni, O. et al: Open Information Extraction from the Web. Comms. of the ACM 51 (12), 68--74 (2008)
11. http://www.cs.washington.edu/research/textrunner/ , “TextRunner Search”, 30 March 2010.
12. http://developer.yahoo.com/search/boss/ , “Yahoo! Search BOSS”, 30 March 2010.
13. http://powerset.com , “Powerset”, 30 March 2010
14. http://www.google.com/squared , “Google Squared”, 30 March 2010.

Appendix: Settings used for Evaluation Runs

English

Cues used in queries to search engine:

is, was, is the, was the, is a, was a, is an, was an, is in, is on, is by, is next to, is near to, is known, is famous, is located, is one of, was built, is made of, is named, was named, is home to, was home to, is used, was used, was completed, was destroyed, was damaged, is the site of, was the site of, was the scene of, was made famous, is the most, is the biggest, is the largest, is the smallest, is the oldest, can be seen from, is popular, is popular with, features, offers, is located by, is located on, is located in, is famous for, is known for, was built by, was built in, was built for, was built to, is open

Regular Expression for Shallow Chunking of Snippets:

```
(^|\.|\,|\;|\:|\?|\!|the|The)\s*ENTITY\s*CUE\s*(.*?)((\.|\,|\;|\
:|\?|\!)|((\b(and)\b|\b(but)\b)))
```

ENTITY and CUE are interpolated at run-time, ‘(.*?)’ captures the Text-Fragment.

Filter words:

I, my, me, mine, you, your, yours, we, us, ours, another, recently, this, also, other, further, must, should, could, sensational, fun, deserves, excellent, amazing, wonderful, miles, kilometres, m, km, minutes, min, mins, hours, hour, probably, actually, possibly

Scoring stop words:

the, of, is, for, a, an, and

Invalid final words:

a, the, those, these, with, by, and, but, which, that, for, like, as

Latvian

Cues used in queries to search engine:

ir, bija, ir pazīstams, ir pazīstama, ir slavens, ir slavena, ir ievērojams, ir ievērojama, atrodas, ir viens no, ir viena no, ir blakus, ir netālu no, tika uzcelts, tika uzcelta, tika celta, tika celts, ir veidots no, ir veidota no, ir izgatavots no, ir izgatavota no, ir nosaukts, ir nosaukta, tika nosaukts, tika nosaukta, bija mājas, tika lietots, tika lietota, tika pabeigts, tika pabeigta, tika sagrauta, tika sagrauts, ir pats, ir pati, ir lielākais, ir lielākā, ir mazākais, ir mazākā, ir vecākais, ir vecākā, ir garākā, ir dziļākā, ir dziļākais, ir augstākais, var redzēt no, ir redzams no, ir redzama no, ir populārs, ir populāra, rāda, atklāj, ir raksturīgs ar, ir raksturīga ar

Regular Expression for Shallow Chunking of Snippets:

```
ENTITY\s*CUE\s*(.*?)(\.|\!|\;|\:|\?|\[)
```

For both languages the maximum snippets returned from search engine for a single query was 20, scoring metric was simple sum, and the score threshold = 3. Note, the word lists were translated for Latvian but did not seem to have much effect.