Interesting Nuggets and Their Impact on Definitional Question Answering

Kian-Wei Kor
Department of Computer Science
School of Computing
National University of Singapore
dkor@comp.nus.edu.sg

Tat-Seng Chua
Department of Computer Science
School of Computing
National University of Singapore
chuats@comp.nus.edu.sg

ABSTRACT

Current approaches to identifying definitional sentences in the context of Question Answering mainly involve the use of linguistic or syntactic patterns to identify informative nuggets. This is insufficient as they do not address the novelty factor that a definitional nugget must also possess. This paper proposes to address the deficiency by building a "Human Interest Model" from external knowledge. It is hoped that such a model will allow the computation of human interest in the sentence with respect to the topic. We compare and contrast our model with current definitional question answering models to show that interestingness plays an important factor in definitional question answering.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models; H.1.2 [**User/Machine Systems**]: Human Factors

General Terms

Algorithms, Human Factors, Experimentation

Keywords

Definitional Question Answering, Human Interest

1. DEFINITIONAL QUESTION ANSWERING

Definitional Question Answering was first introduced to the TExt Retrieval Conference Question Answering Track main task in 2003. The Definition questions, also called Other questions in recent years, are defined as follows. Given a question topic X, the task of a definitional QA system is akin to answering the question *"What is X?"* or *"Who is X?"*. The definitional QA system is to search through a news corpus and return return a set of answers that best describes the question topic. Each answer should be a unique topic-specific nugget that makes up one facet in the definition of the question topic.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

1.1 The Two Aspects of Topic Nuggets

Officially, topic-specific answer nuggets or simply topic nuggets are described as "informative nuggets". Each informative nugget is a sentence fragment that describe some factual information about the topic. Depending on the topic type and domain, this can include topic properties, relationships the topic has with some closely related entity, or events that happened to the topic.

From observation of the answer set for definitional question answering from TREC 2003 to 2005, it seems that a significant number of topic nuggets cannot simply be described as informative nuggets. Rather, these topic nuggets have a trivia-like quality associated with them. Typically, these are out of the ordinary pieces of information about a topic that can pique a human reader's interest. For this reason, we decided to define answer nuggets that can evoke human interest as *"interesting nuggets"*. In essence, interesting nuggets answer the questions *"What is X famous for?"*, *"What defines X?"* or *"What is extraordinary about X?"*.

We now have two very different perspective as to what constitutes an answer to Definition questions. An answer can be some important factual information about the topic or some novel and interesting aspect about the topic. This duality of informativeness and interestingness can be clearly observed in the five vital answer nuggets for a TREC 2005 topic of "George Foreman". Certain answer nuggets are more informative while other nuggets are more interesting in nature.

Informative Nuggets
- Was graduate of Job Corps.
- Became oldest world champion in boxing history.

Interesting Nuggets
- Has lent his name to line of food preparation products.
- Waved American flag after winning 1968 Olympics championship.
- Returned to boxing after 10 yr hiatus.

As an African-American professional heavyweight boxer, an average human reader would find the last three nuggets about George Foreman interesting because boxers do not usually lend their names to food preparation products, nor do boxers retire for 10 years before returning to the ring and become the world's oldest boxing champion. Foreman's waving of the American flag at the Olympics is interesting because the innocent action caused some African-Americans to accuse Foreman of being an Uncle Tom. As seen here, interesting nuggets has some surprise factor or unique quality that makes them interesting to human readers.

1.2 Identifying Interesting Nuggets

Since the original official description for definitions comprise of

identifying informative nuggets, most research has focused entirely on identifying informative nuggets. In this paper, we focus on exploring the properties of interesting nuggets and develop ways of identify such interesting nuggets. A "Human Interest Model" definitional question answering system is developed with emphasis on identifying interesting nuggets in order to evaluate the impact of interesting nuggets on the performance of a definitional question answering system. We further experimented with combining the Human Interest Model with a lexical pattern based definitional question answering system in order to capture both informative and interesting nuggets.

2. RELATED WORK

There are currently two general methods for Definitional Question Answering. The more common method uses a lexical pattern-based approach was first proposed by Blair-Goldensohn et al. [1] and Xu et al. [14]. Both groups predominantly used patterns such as copulas and appositives, as well as manually crafted lexicosyntactic patterns to identify sentences that contain informative nuggets. For example, Xu et al. used 40 manually defined "structured patterns" in their 2003 definitional question answering system. Since then, in an attempt to capture a wider class of informational nuggets, many such systems of increasing complexity has been created. A recent system by Harabagiu et al. [6] created a definitional question answering system that combines the use of 150 manually defined positive and negative patterns, named entity relations and specially crafted information extraction templates for 33 target domains. Here, a musician template may contain lexical patterns that identify information such as the musician's musical style, songs sung by the musician and the band, if any, that the musician belongs to. As one can imagine, this is a knowledge intensive approach that requires an expert linguist to manually define all possible lexical or syntactic patterns required to identify specific types of information.

This process requires a lot of manual labor, expertise and is not scalable. This lead to the development of the soft-pattern approach by Cui et al. [4, 11]. Instead of manually encoding patterns, answers to previous definitional question answering evaluations were converted into generic patterns and a probabilistic model is trained to identify such patterns in sentences. Given a potential answer sentence, the probabilistic model outputs a probability that indicates how likely the sentence matches one or more patterns that the model has seen in training.

Such lexicalosyntactic patterns approach have been shown to be adept at identifying factual informative nuggets such as a person's birthdate, or the name of a company's CEO. However, these patterns are either globally applicable to all topics or to a specific set of entities such as musicians or organizations. This is in direct contrast to interesting nuggets that are highly specific to individual topics and not to a set of entities. For example, the interesting nuggets for George Foreman are specific only George Foreman and no other boxer or human being. Topic specificity or topic relevance is thus an important criteria that helps identify interesting nuggets.

This leads to the exploration of the second relevance-based approach that has been used in definitional question answering. Predominantly, this approach has been used as a backup method for identifying definitional sentences when the primary method of lexicalosyntactic patterns failed to find a sufficient number of informative nuggets [1]. A similar approach has also been used as a baseline system for TREC 2003 [14]. More recently, Chen et al. [3] adapted a bi-gram or bi-term language model for definitional Question Answering.

Generally, the relevance-based approach requires a "definitional corpus" that contain documents highly relevant to the topic. The baseline system in TREC 2003 simply uses the topic words as its definitional corpus. Blair-Goldensohn et al. [1] uses a machine learner to include in the definitonal corpus sentences that are likely to be definitional. Chen et al. [3] collect snippets from Google to build its definitional corpus.

From the definitional corpus, a definitional centroid vector is built or a set of centroid words are selected. This centroid vector or set of centroid words is taken to be highly indicative of the topic. Systems can then use this centroid to identify definitional answers by using a variety of distance metrics to compare against sentences found in the set of retrieved documents for the topic. Blair-Goldensohn et al. [1] uses Cosine similarity to rank sentences by "centrality". Chen et al. [3] builds a bigram language model using the 350 most frequently occurring google snippet terms, described in their paper as an ordered centroid, to estimate the probability that a sentence is similar to the ordered centroid.

As described here, the relevance-based approach is highly specific to individual topics due to its dependence on a topic specific definitional corpus. However if individual sentences are viewed as a document, then relevance-based approaches essentially use the collected topic specific centroid words as a form of document retrieval with automated query expansion to identify strongly relevant sentences. Thus such methods identify relevant sentences and not sentences containing definitional nuggets. Yet, the TREC 2003 baseline system [14] outperformed all but one other system. The bi-term language model [3] is able to report results that are highly competitive to state-of-the-art results using this retrieval-based approach. At TREC 2006, a simple weighted sum of all terms model with terms weighted using solely Google snippets outperformed all other systems by a significant margin [7].

We believe that interesting nuggets often come in the form of trivia, novel or rare facts about the topic that tend to strongly co-occur with direct mention of topic keywords. This may explain why relevance-based method can perform competitively in definitional question answering. However, simply comparing against a single centroid vector or set of centroid words may have over emphasized topic relevance and has only identified interesting definitional nuggets in an indirect manner. Still, relevance based retrieval methods can be used as a starting point in identifying interesting nuggets. We will describe how we expand upon such methods to identify interesting nuggets in the next section.

3. HUMAN INTEREST MODEL

Getting a computer system to identify sentences that a human reader would find interesting is a tall order. However, there are many documents on the world wide web that are contain concise, human written summaries on just about any topic. What's more, these documents are written explicitly for human beings and will contain information about the topic that most human readers would be interested in. Assuming we can identify such relevant documents on the web, we can leverage them to assist in identifying definitional answers to such topics. We can take the assumption that most sentences found within these web documents will contain interesting facets about the topic at hand.

This greatly simplifies the problem to that of finding within the AQUAINT corpus sentences similar to those found in web documents. This approach has been successfully used in several factoid and list Question Answering systems [11] and we feel the use of such an approach for definitional or "Other" question answering is justified. Identifying interesting nuggets requires computing machinery to understand world knowledge and human insight. This is still a very challenging task and the use of human written documents dramatically simplifies the complexity of the task.

In this paper, we report on such an approach by experimenting with a simple word-level edit distance based weighted term comparison algorithm. We use the edit distance algorithm to score the similarity of a pair of sentences, with one sentence coming from web resources and the other sentence selected from the AQUAINT corpus. Through a series of experiments, we will show that even such a simple approach can be very effective at definitional question answering.

3.1 Web Resources

There exists on the internet articles on just about any topic a human can think of. What's more, many such articles are centrally located on several prominent websites, making them an easily accessible source of world knowledge. For our work on identifying interesting nuggets, we focused on finding short one or two page articles on the internet that are highly relevant to our desired topic. Such articles are useful as they contain concise information about the topic. More importantly, the articles are written by humans, for human readers and thus contain the critical human world knowledge that a computer system currently is unable to capture.

We leverage this world knowledge by collecting articles for each topic from the following external resources to build our "Interest Corpus" for each topic.

Wikipedia is a Web-based, free-content encyclopedia written collaboratively by volunteers. This resource has been used by many Question Answering system as a source of knowledge about each topic. We use a snapshot of Wikipedia taken in March 2006 and include the most relevant article in the Interest Corpus.

NewsLibrary is a searchable archive of news articles from over 100 different newspaper agencies. For each topic, we download the 50 most relevant articles and include the title and first paragraph of each article in the Interest Corpus.

Google Snippets are retrieved by issuing the topic as a query to the Google search engine. From the search results, we extracted the top 100 snippets. While Google snippets are not articles, we find that they provide a wide coverage of authorative information about most topics.

Due to their comprehensive coverage of a wide variety of topics, the above resources form the bulk of our Interest Corpus. We also extracted documents from other resources. However, as these resources are more specific in nature, we do not always get any single relevant document. These resources are listed below.

Biography.com is the website for the Biography television cable channel. The channel's website contains searchable biographies on over 25,000 notable people. If the topic is a person and we can find a relevant biography on the person, we include it it in our Interest Corpus.

Bartleby.com contains a searchable copy of several resources including the Columbia Encyclopedia, the World Factbook, and several English dictionaries.

s9.com is a biography dictionary on over 33,000 notable people. Like Biography.com, we include the most relevant biography we can find in the Interest Corpus.

Google Definitions Google search engine offers a feature called "Definitions" that provides the definition for a query, if it has one. We use this feature and extract whatever definitions the Google search engine has found for each topic into the Interest Corpus.

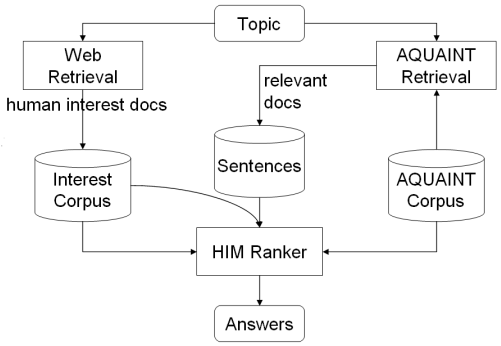

Figure 1: Human Interest Model Architecture.

WordNet WordNet is an well-known electronic semantic lexicon for the English language. Besides grouping English words into sets of synonyms called synsets, it also provide a short definition on the meaning of words found in each synset. We add this short definition, if there is one, into our Interest Corpus.

We have two major uses for this topic specific Interest Corpus, as a source of sentences containing interesting nuggets and as a unigram language model of topic terms, I.

3.2 Multiple Interesting Centroids

We have seen that interesting nuggets are highly specific to a topic. Relevance-based approaches such as the bigram language model used by Chen et al. [3] are focused on identifying highly relevant sentences and pick up definitional answer nuggets as an indirect consequence. We believe that the use of only a single collection of centroid words has over-emphasized topic relevance and choose instead to use multiple "centroids".

Since sentences in the Interest Corpus of articles we collected from the internet are likely to contain nuggets that are of interest to human readers, we can essentially use each sentence as "pseudo-centroids". Each sentence in the Interest Corpus essentially raises a different aspect of the topic for consideration as a sentence of interest to human readers. By performing a pairwise sentence comparison between sentences in the Interest Corpus and candidate sentences retrieved from the AQUAINT corpus, we increase the number of sentence comparisons from $O(n)$ to $O(nm)$. Here, n is the number of potential candidate sentences and m is the number of sentences in the Interest Corpus. In return, we obtain a diverse ranked list of answers that are individually similar to various sentences found in the topic's Interest Corpus. An answer can only be highly ranked if it is strongly similar to a sentence in the Interest Corpus, and is also strongly relevant to the topic.

3.3 Implementation

Figure 1 shows the system architecture for the proposed Human Interest-based definitional QA system.

The AQUAINT Retrieval module shown in Figure 1 reuses a document retrieval module of a current Factoid and List Question Answering system we have implemented. Given a set of words describing the topic, the AQUAINT Retrieval module does query expansion using Google and searches an index of AQUAINT documents to retrieve the 800 most relevant documents for consideration.

The Web Retrieval module on the other hand, searches the online

resources described in Section 3.1 for "interesting" documents in order to populate the Interest Corpus.

The HIM Ranker, or Human Interest Model Ranking module, is the implementation of what is described in this paper. The module first builds the unigram language model, I, from the collected web documents. This language model will be used to weight the importance of terms within sentences. Next, a sentence chunker is used to segment all 800 retrieved documents into individual sentences. Each of these sentences can be a potential answer sentence that will be independently ranked by interestingness. We rank sentences by interestingness using sentences from both the Interest Corpus of external documents as well as the unigram language model we built earlier which we use to weight terms.

A candidate sentence in our top 800 relevant AQUAINT documents is considered interesting if it is highly similar in content to a sentence found in our collection of external web-documents. To achieve this, we perform a pairwise similarity comparison between a candidate sentence and sentences in our external documents using a weighted-term edit distance algorithm. Term weights are used to adjust the relative importance of each unique term found in the Interest Corpus. When both sentences share the same term, the similarity score is incremented by the two times the term's weight and every dissimilar term decrements the similarity score by the dissimilar term's weight.

We choose the highest achieved similarity score for a candidate sentence as the Human Interest Model score for the candidate sentence. In this manner, every candidate sentence is ranked by interestingness. Finally, to obtain the answer set, we select the top 12 highest ranked and non redundant sentences as definitional answers for the topic.

4. INITIAL EXPERIMENTS

The Human Interest-based system described in the previous section is designed to identify only interesting nuggets and not informative nuggets. Thus, it can be described as a handicapped system that only deals with half the problem in definitional question answering. This is done in order to explore how interestingness plays a factor in definitional answers. In order to compare and contrast the differences between informative and interesting nuggets, we also implemented the soft-pattern bigram model proposed by Cui et al. [4, 11]. In order to ensure comparable results, both systems are provided identical input data. Since both system require the use of external resources, they are both provided the same web articles retrieved by our Web Retrieval module. Both systems also rank the same same set of candidate sentences in the form of 800 most relevant documents as retrieved by our AQUAINT Retrieval module.

For the experiments, we used the TREC 2004 question set to tune any system parameters and use the TREC 2005 question sets to test the both systems. Both systems are evaluated the results using the standard scoring methodology for TREC definitions. TREC provides a list of vital and okay nuggets for each question topic. Every question is scored on nugget recall (NR) and nugget precision (NP) and a single final score is computed using F-Measure (see equation 1) with $\beta = 3$ to emphasize nugget recall. Here, NR is the number of vital nuggets returned divided by total number of vital nuggets while NP is computed using a minimum allowed character length function defined in [12]. The evaluation is automatically conducted using Pourpre v1.0c [10].

$$FScore = \frac{\beta^2 * NP * NR}{(\beta^2 + 1)NP + NR} \quad (1)$$

System	F3-Score
Best TREC 2005 System	0.2480
Soft-Pattern (SP)	0.2872
Human Interest Model (HIM)	0.3031

Table 1: Performance on TREC 2005 Question Set

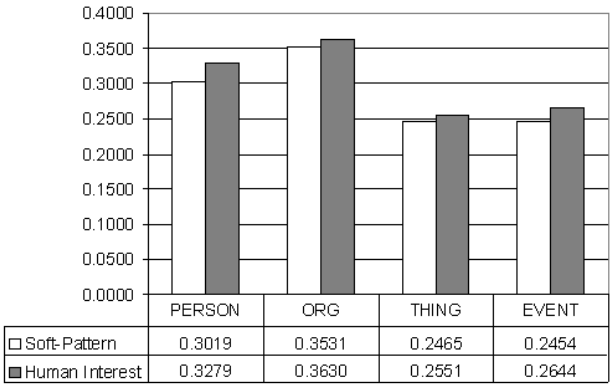

Figure 2: Performance by entity types.

4.1 Informativeness vs Interestingness

Our first experiment compares the performance of solely identifying interesting nuggets against solely identifying informative nuggets. We compare the results attained by the Human Interest Model that only identify interesting nuggets with the results of the syntactic pattern finding Soft-Pattern model as well as the result of the top performing definitional system in TREC 2005 [13]. Table 1 shows the F3 score the three systems for the TREC 2005 question set.

The Human Interest Model clearly outperform both soft pattern and the best TREC 2005 system with a F3 score of 0.303. The result is also comparable with the result of a human manual run, which attained a F3 score of 0.299 on the same question set [9]. This result is confirmation that interesting nuggets does indeed play a significant role in picking up definitional answers, and may be more vital than using information finding lexical patterns.

In order to get a better perspective of how well the Human Interest Model performs for different types of topics, we manually divided the TREC 2005 topics into four broad categories of PERSON, ORGANIZATION, THING and EVENT as listed in Table 3. These categories conform to TREC's general division of question topics into 4 main entity types [13]. The performance of Human Interest Model and Soft Pattern Bigram Model for each entity type can be seen in Figure 2. Both systems exhibit consistent behavior across entity types, with the best performance coming from PERSON and ORGANIZATION topics and the worst performance from THING and EVENT topics. This can mainly be attributed to our selection of web-based resources for the definitional corpus used by both system. In general, it is harder to locate a single web article that describes an event or a general object. However given the same set of web-based information, the Human Interest Model consistently outperforms the soft-pattern model for all four entity types. This suggests that the Human Interest Model is better able to leverage the information found in web resources to identify definitional answers.

5. REFINEMENTS

Encouraged by the initial experimental results, we explored two further optimization of the basic algorithm.

5.1 Weighting Interesting Terms

The word trivia refer to tidbits of unimportant or uncommon information. As we have noted, interesting nuggets often has a trivia-like quality that makes them of interest to human beings. From this description of interesting nuggets and trivia, we hypothesize that interesting nuggets are likely to occur rarely in a text corpora.

There is a possibility that some low-frequency terms may actually be important in identifying interesting nuggets. A standard unigram language model would not capture these low-frequency terms as important terms. To explore this possibility, we experimented with three different term weighting schemes that can provide more weight to certain low-frequency terms. The weighting schemes we considered include commonly used TFIDF, as well as information theoretic Kullback-Leiber divergence and Jensen-Shannon divergence [8].

TFIDF, or Term Frequency $\times$ Inverse Document Frequency, is a standard Information Retrieval weighting scheme that balances the importance of a term in a document and in a corpus. For our experiments, we compute the weight of each term as $tf \times log(\frac{N}{n_t})$, where tf is the term frequency, n_t is the number of sentences in the Interest Corpus having the term and N is the total number of sentences in the Interest Corpus.

Kullback-Leibler Divergence (Equation 2) is also called KL Divergence or relative entropy, can be viewed as measuring the dissimilarity between two probability distributions. Here, we treat the AQUAINT corpus as a unigram language model of general English [15], A, and the Interest Corpus as a unigram language model consisting of topic specific terms and general English terms, I. General English words are likely to have similar distributions in both language models I and A. Thus using KL Divergence as a term weighting scheme will cause strong weights to be given to topic-specific terms because their distribution in the Interest Corpus they occur significantly more often or less often than in general English. In this way, high frequency centroid terms as well as low frequency rare but topic-specific terms are both identified and highly weighted using KL Divergence.

$$D_{KL}(I \parallel A) = \sum_t I(t) log \frac{I(t)}{A(t)} \quad (2)$$

Due to the power law distribution of terms in natural language, there are only a small number of very frequent terms and a large number of rare terms in both I and A. While the common terms in English consist of stop words, the common terms in the topic specific corpus, I, consist of both stop words and relevant topic words. These high frequency topic specific words occur very much more frequently in I than in A. As a result, we found that KL Divergence has a bias towards highly frequent topic terms as we are measuring direct dissimilarity against a model of general English where such topic terms are very rare. For this reason, we explored another divergence measure as a possible term weighting scheme.

Jensen-Shannon Divergence or JS Divergence extends upon KL Divergence as seen in Equation 3. As with KL Divergence, we also use JS divergence to measure the dissimilarity between our two language models, I and A.

$$D_{JS}(I \parallel A) = \frac{1}{2}\left[D_{KL}\left(I \| \frac{I+A}{2}\right) + D_{KL}\left(A \parallel \frac{I+A}{2}\right)\right] \quad (3)$$

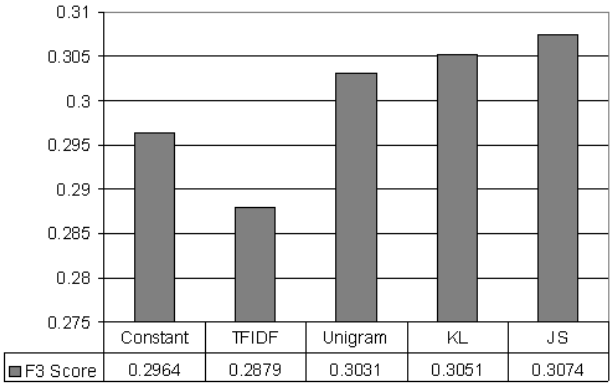

Figure 3: Performance by various term weighting schemes on the Human Interest Model.

However, JS Divergence has additional properties[1] of being symmetric and non-negative as seen in Equation 4. The symmetric property gives a more balanced measure of dissimilarity and avoids the bias that KL divergence has.

$$D_{JS}(I \| A) = D_{JS}(A \| I) = \begin{cases} 0 & I = A \\ > 0 & I <> A \end{cases} \quad (4)$$

We conducted another experiment, substituting the unigram languge model weighting scheme we used in the initial experiments with the three term weighting schemes described above. As lower bound reference, we included a term weighting scheme consisting of a constant 1 for all terms. Figure 3 show the result of applying the five different term weighting schemes on the Human Interest Model. TFIDF performed the worst as we had anticipated. The reason is that most terms only appear once within each sentence, resulting in a term frequency of 1 for most terms. This causes the IDF component to be the main factor in scoring sentences. As we are computing the Inverse Document Frequency for terms in the Interest Corpus collected from web resources, IDF heavily down-weights highly frequency topic terms and relevant terms. This results in TFIDF favoring all low frequency terms over high frequency terms in the Interest Corpus. Despite this, the TFIDF weighting scheme only scored a slight 0.0085 lower than our lower bound reference of constant weights. We view this as a positive indication that low frequency terms can indeed be useful in finding interesting nuggets.

Both KL and JS divergence performed marginally better than the uniform language model probabilistic scheme that we used in our initial experiments. From inspection of the weighted list of terms, we observed that while low frequency relevant terms were boosted in strength, high frequency relevant terms still dominate the top of the weighted term list. Only a handful of low frequency terms were weighted as strongly as topic keywords and combined with their low frequency, may have limited the impact of re-weighting such terms. However we feel that despite this, Jensen-Shannon divergence does provide a small but measurable increase in the performance of our Human Interest Model.

[1]JS divergence also has the property of being bounded, allowing the results to be treated as a probability if required. However, the bounded property is not required here as we are only treating the divergence computed by JS divergence as term weights

5.2 Selecting Web Resources

In one of our initial experiments, we observed that the quality of web resources included in the Interest Corpus may have a direct impact on the results we obtain. We wanted to determine what impact the choice of web resources have on the performance of our Human Interest Model. For this reason, we split our collection of web resources into four major groups listed here:

N - News: Title and first paragraph of the top 50 most relevant articles found in NewsLibrary.

W - Wikipedia: Text from the most relevant article found in Wikipedia.

S - Snippets: Snippets extracted from the top 100 most relevant links after querying Google.

M - Miscellaneous sources: Combination of content (when available) from secondary sources including biography.com, s9.com, bartleby.com articles, Google definitions and WordNet definitions.

We conducted a gamut of runs on the TREC 2005 question set using all possible combinations of the above four groups of web resources to identify the best possible combination. All runs were conducted on Human Interest Model using JS divergence as term weighting scheme. The runs were sorted in descending F3-Score and the top 3 best performing runs for each entity class are listed in Table 2 together with earlier reported F3-scores from Figure 2 as a baseline reference. A consistent trend can be observed for each entity class.

For PERSON and EVENT topics, NewsLibrary articles are the main source of interesting nuggets with Google snippets and miscellaneous articles offering additional supporting evidence. This seem intuitive for events as newspapers predominantly focus on reporting breaking newsworthy events and are thus excellent sources of interesting nuggets. We had expected Wikipedia rather than news articles to be a better source of interesting facts about people and were surprised to discover that news articles outperformed Wikipedia. We believe that the reason is because the people selected as topics thus far have been celebrities or well known public figures. Human readers are likely to be interested in news events that spotlight these personalities.

Conversely for ORGANIZATION and THING topics, the best source of interesting nuggets come from Wikipedia's most relevant article on the topic with Google snippets again providing additional information for organizations.

With an oracle that can classify topics by entity class with 100% accuracy and by using the best web resources for each entity class as shown in Table 2, we can attain a F3-Score of 0.3158.

Rank	PERSON	ORG	THING	EVENT
Baseline	Unigram Weighting Scheme, N+W+S+M			
	0.3279	0.3630	0.2551	0.2644
1	N+S+M	W+S	W+M	N+M
	0.3584	0.3709	0.2688	0.2905
2	N+S	N+W+S	W+S+M	N+S+M
	0.3469	0.3702	0.2665	0.2745
3	N+M	N+W+S+M	W+S	N+S
	0.3431	0.3680	0.2616	0.2690

Table 2: Top 3 runs using different web resources for each entity class

6. UNIFYING INFORMATIVENESS WITH INTERESTINGNESS

We have thus far been comparing the Human Interest Model against the Soft-Pattern model in order to understand the differences between interesting and informative nuggets. However from the perspective of a human reader, both informative and interesting nuggets are useful and definitional. Informative nuggets present a general overview of the topic while interesting nuggets give readers added depth and insight by providing novel and unique aspects about the topic. We believe that a good definitional question answering system should provide the reader with a combined mixture of both nugget types as a definitional answer set.

We now have two very different "experts" at identifying definitions. The Soft Pattern Bigram Model proposed by Cui et al. is an expert in identifying informative nuggets. The Human Interest Model we have described in this paper on the other hand is an expert in finding interesting nuggets. We had initially hoped to unify the two separate definitional question answering systems by applying an ensemble learning method [5] such as voting or boosting in order to attain a good mixture of informative and interesting nuggets in our answer set. However, none of the ensemble learning methods we attempted could outperform our Human Interest Model.

The reason is that both systems are picking up very different sentences as definitional answers. In essence, our two experts are disagreeing on which sentences are definitional. In the top 10 sentences from both systems, only 4.4% of these sentences appeared in both answer sets. The remaining answers were completely different. Even when we examined the top 500 sentences generated by both systems, the agreement rate was still an extremely low 5.3%. Yet, despite the low agreement rate between both systems, each individual system is still able to attain a relatively high F3 score.

There is a distinct possibility that each system may be selecting different sentences with different syntactic structures but actually have the same or similar semantic content. This could result in both systems having the same nuggets marked as correct even though the source answer sentences are structurally different. Unfortunately, we are unable to automatically verify this as the evaluation software we are using does not report correctly identified answer nuggets.

To verify if both systems are selecting the same answer nuggets, we randomly selected a subset of 10 topics from the TREC 2005 question set and manually identified correct answer nuggets (as defined by TREC accessors) from both systems. When we compared the answer nuggets found by both system for this subset of topics, we found that the nugget agreement rate between both systems was 16.6%. While the nugget agreement rate is higher than the sentence agreement rate, both systems are generally still picking up different answer nuggets. We view this as further indication that definitions are indeed made up of a mixture of informative and interesting nuggets. It is also indication that in general, interesting and informative nuggets are quite different in nature.

There are thus rational reasons and practical motivation in unifying answers from both the pattern based and corpus based approaches. However, the differences between the two systems also cause issues when we attempt to combine both answer sets. Currently, the best approach we found for combining both answer sets is to merge and re-rank both answer sets with boosting agreements.

We first normalize the top 1,000 ranked sentences from each system, to obtain the Normalized Human Interest Model score, $him(s)$, and the Normalized Soft Pattern Bigram Model score,

$sp(s)$, for every unique sentence, s. For each sentence, the two separate scores for are then unified into a single score using Equation 5. When only one system believes that the sentence is definitional, we simply retain that system's normalized score as the unified score. When both systems agree agree that the sentence is definitional, the sentence's score is boosted by the degree of agreement between between both systems.

$$Score(s) = max(s_{him}, s_{sp})^{1-min(s_{him}, s_{sp})} \quad (5)$$

In order to maintain a diverse set of answers as well as to ensure that similar sentences are not given similar ranking, we further re-rank our combined list of answers using Maximal Marginal Relevance or MMR [2]. Using the approach described here, we achieve a F3 score of 0.3081. This score is equivalent to the initial Human Interest Model score of 0.3031 but fails to outperform the optimized Human Interest Model model.

7. CONCLUSION

This paper has presented a novel perspective for answering definitional questions through the identification of interesting nuggets. Interesting nuggets are uncommon pieces of information about the topic that can evoke a human reader's curiosity. The notion of an "average human reader" is an important consideration in our approach. This is very different from the lexico-syntactic pattern approach where the context of a human reader is not even considered when finding answers for definitional question answering.

Using this perspective, we have shown that using a combination of a carefully selected external corpus, matching against multiple centroids and taking into consideration rare but highly topic specific terms, we can build a definitional question answering module that is more focused on identifying nuggets that are of interest to human beings. Experimental results has shown this approach can significantly outperform state-of-the-art definitional question answering systems.

We further showed that at least two different types of answer nuggets are required to form a more thorough set of definitional answers. What seems to be a good set of definition answers is some general information that provides a quick informative overview mixed together with some novel or interesting aspects about the topic. Thus we feel that a good definitional question answering system would need to pick up both informative and interesting nugget types in order to provide a complete definitional coverage on all important aspects of the topic. While we have attempted to build such a system by combining our proposed Human Interest Model with Cui et al.'s Soft Pattern Bigram Model, the inherent differences between both types of nuggets seemingly caused by the low agreement rates between both models have made this a difficult task. Indeed, this is natural as the two models have been designed to identify two very different types of definition answers using very different types of features. As a result, we are currently only able to achieve a hybrid system that has the same level of performance as our proposed Human Interest Model.

We approached the problem of definitional question answering from a novel perspective, with the notion that interest factor plays a role in identifying definitional answers. Although the methods we used are simple, they have been shown experimentally to be effective. Our approach may also provide some insight into a few anomalies in past definitional question answering's trials. For instance, the top definitional system at the recent TREC 2006 evaluation was able to significantly outperform all other systems using relatively simple unigram probabilities extracted from Google snippets. We suspect the main contributor to the system's performance is Google's PageRank algorithm, which mainly consider the number of linkages, has an indirect effect of ranking web documents by the degree of human interest.

In our future work, we seek to further improve on the combined system by incorporating more evidence in support of correct definitional answers or to filter away obviously wrong answers.

Entity Type	Topics
ORGANIZATION	DePauw University, Merck & Co., Norwegian Cruise Lines (NCL), United Parcel Service (UPS), Little League Baseball, Cliffs Notes, American Legion, Sony Pictures Entertainment (SPE), Telefonica of Spain, Lions Club International, AMWAY, McDonald's Corporation, Harley-Davidson, U.S. Naval Academy, OPEC, NATO, International Bureau of Universal Postal Union (UPU), Organization of Islamic Conference (OIC), PBGC
PERSON	Bing Crosby, George Foreman, Akira Kurosawa, Sani Abacha, Enrico Fermi, Arnold Palmer, Woody Guthrie, Sammy Sosa, Michael Weiss, Paul Newman, Jesse Ventura, Rose Crumb, Rachel Carson, Paul Revere, Vicente Fox, Rocky Marciano, Enrico Caruso, Pope Pius XII, Kim Jong Il
THING	F16, Bollywood, Viagra, Howdy Doody Show, Louvre Museum, meteorites, Virginia wine, Counting Crows, Boston Big Dig, Chunnel, Longwood Gardens, Camp David, kudzu, U.S. Medal of Honor, tsunami, genome, Food-for-Oil Agreement, Shiite, Kinmen Island
EVENT	Russian submarine Kursk sinks, Miss Universe 2000 crowned, Port Arthur Massacre, France wins World Cup in soccer, Plane clips cable wires in Italian resort, Kip Kinkel school shooting, Crash of EgyptAir Flight 990, Preakness 1998, first 2000 Bush-Gore presidential debate , 1998 indictment and trial of Susan McDougal, return of Hong Kong to Chinese sovereignty, 1998 Nagano Olympic Games, Super Bowl XXXIV, 1999 North American International Auto Show, 1980 Mount St. Helens eruption, 1998 Baseball World Series, Hindenburg disaster, Hurricane Mitch

Table 3: TREC 2005 Topics Grouped by Entity Type

8. REFERENCES

[1] S. Blair-Goldensohn, K. R. McKeown, and A. H. Schlaikjer. A hybrid approach for qa track definitional questions. In *TREC '03: Proceedings of the 12th Text REtrieval Conference*, Gaithersburg, Maryland, 2003.

[2] J. G. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Research and Development in Information Retrieval*, pages 335–336, 1998.

[3] Y. Chen, M. Zhou, and S. Wang. Reranking answers for definitional qa using language modeling. In *Proceedings of*

the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 1081–1088, Sydney, Australia, July 2006. Association for Computational Linguistics.

[4] H. Cui, M.-Y. Kan, and T.-S. Chua. Generic soft pattern models for definitional question answering. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 384–391, New York, NY, USA, 2005. ACM Press.

[5] T. G. Dietterich. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15, 2000.

[6] S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, A. Hickl, and P. Wang. Employing two question answering systems at trec 2005. In *TREC '05: Proceedings of the 14th Text REtrieval Conference*, Gaithersburg, Maryland, 2005.

[7] M. Kaisser, S. Scheible, and B. Webber. Experiments at the university of edinburgh for the trec 2006 qa track. In *TREC '06 Notebook: Proceedings of the 14th Text REtrieval Conference*, Gaithersburg, Maryland, 2006. National Institute of Standards and Technology.

[8] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145 – 151, Jan 1991.

[9] J. Lin, E. Abels, D. Demner-Fushman, D. W. Oard, P. Wu, and Y. Wu. A menagerie of tracks at maryland: Hard, enterprise, qa, and genomics, oh my! In *TREC '05: Proceedings of the 14th Text REtrieval Conference*, Gaithersburg, Maryland, 2005.

[10] J. Lin and D. Demner-Fushman. Automatically evaluating answers to definition questions. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 931–938, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[11] R. Sun, J. Jiang, Y. F. Tan, H. Cui, T.-S. Chua, and M.-Y. Kan. Using syntactic and semantic relation analysis in question answering. In *TREC '05: Proceedings of the 14th Text REtrieval Conference*, Gaithersburg, Maryland, 2005.

[12] E. M. Voorhees. Overview of the trec 2003 question answering track. In *Text REtrieval Conference 2003*, Gaithersburg, Maryland, 2003. National Institute of Standards and Technology.

[13] E. M. Voorhees. Overview of the trec 2005 question answering track. In *TREC '05: Proceedings of the 14th Text REtrieval Conference*, Gaithersburg, Maryland, 2005. National Institute of Standards and Technology.

[14] J. Xu, A. Licuanan, and R. Weischedel. TREC 2003 QA at BBN: Answering definitional questions. In *TREC '03: Proceedings of the 12th Text REtrieval Conference*, Gaithersburg, Maryland, 2003.

[15] D. Zhang and W. S. Lee. A language modeling approach to passage question answering. In *TREC '03: Proceedings of the 12th Text REtrieval Conference*, Gaithersburg, Maryland, 2003.