Web Searching with Multiple Correct Answers

Steven O'Hara
University of Texas San Antonio
One UTSA Circle
San Antonio, TX 78249
(1) 210-458-4436
steve@oharasteve.com

Tom Bylander
University of Texas San Antonio
One UTSA Circle
San Antonio, TX 78249
(1) 210-458-5693
bylander@cs.utsa.edu

ABSTRACT
Most web search engines today are geared towards providing a list of relevant websites, along with snippets of text from each website that are relevant to the user's search text. Some of them may also provide specific answers to the user's question. This paper explores techniques for combining many candidate answers into a small set of answers, when it is likely that there is more than one correct answer. We describe and test new algorithms that collect and consolidate candidate answers from many different websites using paired support, reinforcing the ranking factor of those candidate answers that co-occur as pairs across multiple website domains.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *question-answering (fact retrieval) systems.*

General Terms
Algorithms, Reliability, Verification.

Keywords
Text Retrieval, Question Answering, Intelligent Search, Multiple Correct Answers.

1. INTRODUCTION
Consider the following questions:

q1: *What is the population of New York City?*
q2: *What is the area of Canada?*
q3: *When did Elizabeth Taylor get married?*
q4: *What is the capital of Bolivia?*
q5: *Who coaches the Hoyas men's basketball team?*

In each case, the question appears to imply a single correct answer, and each question is unambiguous in and of itself. But each has more than one correct answer. Consider q1, does the user want the metropolitan statistical area, the seven boroughs, or some other physical boundary? In q2, do they want to include interior rivers and lakes? In q3, she was married eight different times, to seven different men. In q4, the de facto capital is La Paz while the official capital is Sucre. In q5, there is a main coach, but also several assistant coaches.

*WIMS'*14, June 2-4, 2014 Thessaloniki, Greece

http://dx.doi.org/10.1145/2611040.2611071

Answers to questions like these are generally referred to as factoids [1], which are <object, attribute, value> triples. For our purposes, we use <object, attribute, values> instead. See [2] for some related research when there is a single correct answer.

Correct answers often depend on time as well. In some sense, web searches have an implied time reference to the current time. In q1, the population of New York City likely changes slightly many times every day.

Our hypothesis is that the search performance will be increased by (1) allowing similar values to support each other, and by (2) boosting pairs of values that co-occur in multiple domains. Our experiments show that different ways of applying these two criteria result in a significant improvement over the baseline case.

2. RELATED RESEARCH
There are other ways of observing multiple correct answers that have been explored elsewhere.

2.1 Ambiguous Questions
Does the following question refer to Paris, France or Paris, Texas?

q6: *What is the population of Paris?*

This type of ambiguity is in the question, which is not part of this research. If a model of the user is available, or if the location of the user is known, then perhaps the ambiguity could be resolved. Dealing with ambiguity [3] and user models [4] are classic problems in natural language processing.

2.2 Real World Measurements
Many questions reference a real world measurement; for example:

q7: *What is the diameter of Pluto?*

Most real world measurements are imprecise for a variety of reasons. Usually the accuracy of the measurements will improve over time. Consider q7, the diameter of Pluto, where the precision of the measurements improves over time as our astronomy technology improves.[1]

2.3 Corroborating Answers
Wu and Marian [2] "*focus more specifically on those queries for which there is no agreed upon correct answer*", using the example of gas mileage in cars.

They consider the number, importance, and similarity of the web sources reporting each answer, as well as the importance of the answer within the source. Like our research, they attempt to provide the best consensus answer rather than just a collection of snippets that may contain candidate answers.

In their results, they were able to demonstrate that their system was effective and significantly faster than a user clicking

[1] wikipedia: Pluto#Mass_and_size

through multiple websites to attempt to identify a consensus answer.

3. NUMERIC METHODOLOGY

In this section, we focus on questions with (potentially) multiple correct numeric answers, all positive. Two independent domains are considered: populations of U.S. cities and areas of countries. Both are numeric and both can have multiple correct answers.

In the development phase, a small sample of questions (10%) were used to develop our methods. All the questions (100%) were used for the results presented in this paper. Target answers were collected in advance.

3.1 Precision and Recall

When a question has multiple correct answers, there is often no definitive way to ascertain what those correct answers are. It may be a matter of opinion, or a difference in definitions, or a variety of other reasons. Hence, it is challenging to come up with a method of validating the results of this research.

Generally speaking, precision is a measure of how often a candidate answer is correct, and recall is a measure of how many of the actual answers were found. More precisely, if

> Targets ≡ set of all relevant documents, and
> Answers ≡ set of all retrieved documents, then
> Precision ≡ | Targets ∩ Answers |/| Answers |, and
> Recall ≡ | Targets ∩ Answers |/| Targets |.

The measure of precision is still applicable when there are multiple correct answers. However, recall does not apply directly, because the target answers are not always equally important, and there may be a continuum of correct answers. In this paper, we use the term "Rightiness" (see Section 3.6) as our measure of precision.

3.2 Questions and Candidate Answers

Each question was sent to Google's search engine, and up to 500 results were returned. Only the two or three line snippets returned were used in the analysis, not the referenced website.

Each snippet was analyzed looking for very specific patterns that might be candidate answers. These patterns are all simple regular expressions and do not involve parsing the snippet text.

Candidate answers are filtered for obviously wrong answers (out of range), but it is fully expected that some incorrect answers will be found. Our hypothesis is that correct answers will appear in many snippets, and incorrect answers will only appear in a few of the websites. Our expectation is that the results will have a bimodal distribution when there are two correct answers and, more generally, a multimodal distribution where the number of peaks corresponds to the number of correct answers.

3.3 Baseline Case

We compare our results against a baseline case. Each candidate answer (X_i) is assigned a weight, which is computed as:

$$Gap_{i,j} \equiv Max(X_i, X_j) / 100, \text{ for all } i \neq j$$
$$Same_{i,j} \equiv 1 \text{ if } |X_i - X_j| < Gap_{i,j}; \text{ 0 otherwise}$$
$$Weight_i = Sum(Same_{i,j}) \text{ for all } i, j$$

Essentially, this accumulates values that are within 1% of each other. These values are then sorted by weight, and the top ranked values are chosen as the computed answers.

3.4 Calculating Support

The concepts of "Support" and "Rightiness" (discussed in Section 3.6) were developed in [5] and [6].

Given two numeric, positive, answers X and Y, already converted to the same units, and each with its own positive delta interval ($\Delta X > 0$ and $\Delta Y > 0$), define their ranges:

$$X_{LO} = X - \Delta X \qquad Y_{LO} = Y - \Delta Y$$
$$X_{HI} = X + \Delta X \qquad Y_{HI} = Y + \Delta Y$$

To determine the amount of Support X gives to Y, we use:

$$(\min(X_{HI}, Y_{HI}) - \max(X_{LO}, Y_{LO})) \div (X_{HI} - X_{LO})$$

For example, suppose X is 35 ± 25 and Y is 90 ± 50, then Support is calculated as:

$$Support(X,Y) = (60-40) \div (60-10) = 40\%$$
$$Support(Y,X) = (60-40) \div (140-40) = 20\%$$

Effectively, 40% of X's interval is in Y, and 20% of Y's interval is in X.

For example, consider again q7, "What is the diameter of Pluto?" Five of the 25 candidate answers were over 1,500 miles and are not visible in Figure 1. The remaining candidate answers are sorted by their interval size. The bottom axis is just the candidate answer index; the left axis is miles.

In our analysis, the implied interval size is based on the apparent representation of the candidate answer. E.g., an answer of 1,400 has an implied interval of ± 50, while an answer of 1,413 has an implied interval of only ± 0.5. See our prior work [5] for additional analysis.

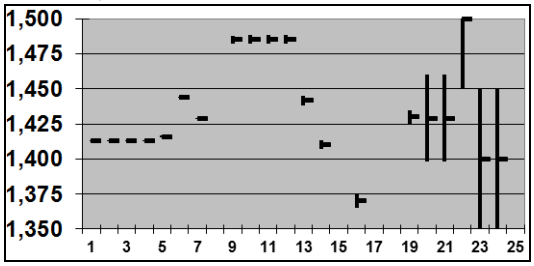

Figure 1: Diameter of Pluto Candidate Answers.

It is easy to see that there are a variety of answers, and there is also a wide range of implied interval sizes. The two candidate answers of 1,400 miles receive the most support from other candidate answers. The four candidate answers of 2,274 kilometers (≈1,413 miles) give the most support to the other candidate answers.

Our ranking prefers the specific answer of 2,274 kilometers to the general answer of 1,400 ± 50 miles in the above example. The mean diameter of Pluto is 2,306 km, which is about 1,433 miles.[2]

3.5 Calculating Paired Support

This is the centerpiece of this research. Instead of just relying on all the individual candidate answers, we consider how pairs of these answers are distributed across multiple web domains. We collect all the candidate answers for each domain, looking only at the domain name and ignoring everything else in the URL.

The Paired Support algorithm in Figure 2 finds all co-occurring pairs of values in different domains, and chooses the largest co-occurring paired support value. It uses the Support function described in Section 3.4.

[2] wikipedia: Pluto

```
For each domain, call it X {
  For each pair of answers in X: (Xa, Xb) {
    Set MaxCPS = 0
    For each other domain Y ≠ X {
      For each pair of answers in Y: (Ya, Yb) {
        // CPS is Co-occurring Paired Support
        CPS =  Support(Xa,Ya) + Support(Xa,Yb) +
               Support(Xb,Ya) + Support(Xb,Yb)
        If CPS > MaxCPS, set MaxCPS = CPS
      }
    }
    Set Paired Support for (Xa, Xb) to MaxCPS.
  }
}
```

Figure 2: Paired Support Algorithm.

3.6 Calculating Rightiness

Let C be a candidate answer, and let T be a target answer (both positive). We define:

$$R(C,T) = 1 - (\min(C,T) / \max(C,T)) + 0.001$$
$$Rightiness(C,T) = \max(0, -\log_{10}(R(C,T)) / 3)$$

Rightiness roughly measures how many significant digits match, up to 3 digits. It will result in about 0.3 if only the first significant digit matches, about 0.6 if the first two significant digits match, and about 0.9 if the first three significant digits match. A result of 1.0 is an exact match.

For calculating the Rightiness of pairs of answers, we use the average of the two individual Rightiness measures.

3.7 Ranking and Scoring Candidate Answers

We use a variation of the Inverse Reciprocal Rank (IRR) algorithm [7], adjusted to handle multiple target answers. We collect all the candidate answers and rank them by the amount of Support they give the other candidate answers. Since there can be multiple target answers, we calculate the precision score as:

$$\text{for each target answer } T_i:$$
$$S_i = \text{Max}_{\{\text{for each k in C}\}} (Rightiness(C_k,T_i) / k)$$
$$Precision = Average(S_i)$$

Each S_i is calculated using ordinary IRR, where the Max function iterates through all the Candidate answers (C_k), starting with k = 1 for the first candidate answer, k = 2 for the second, etc. The final precision score is the average of the S_i values.

4. CITY POPULATION RESULTS

This section is motivated by this question:

q8: *What is the population of Tampa, Florida?*

For this case, 149 U.S. cities are being studied. In most cities, there are at least two measurements: the legal city limits and the metropolitan statistical area. Based on [8], we used 5% of the population count as the delta for calculating Support (see Section 3.4). E.g., we interpret a candidate answer of 112,205 as the range 106,594 to 117,815.

4.1 City Population Target Answers

Each data element has a city name, a metropolitan statistical area population[3] and a city population[4]. These values clearly change over time, and many websites will report historical counts, population changes, or demographics.

Table 1 lists the first three target answers (from Wikipedia).

Table 1: Population Target Answers.

City, State	City	MSA
Abilene, Texas	118,117	166,416
Akron, Ohio	198,402	701,456
Albuquerque, New Mexico	552,804	898,642

4.2 City Population Search Results

For each city, two queries are sent to Google search, such as:

```
What is the Population of Akron, Ohio
      thousand million num=500
What is the Population of Akron, Ohio
      num=500 as_nlo=100000 as_nhi=6926535
```

The first query requests up to 500 websites with the words thousand or million in them. The second is a Google advanced search that finds numbers in the range 100,000 to 60% of the population of the state (Ohio was 11,544,225[5]). State populations were supplied[6] in advance and used in the calculations.

Table 2 lists a few of the candidate answers (answers from Wikipedia were removed).

Notice the entry with 200,000 "Diamond" brand lamps. Since this study is based on reasoning without natural language processing, we expect a significant number of candidate answers to be unrelated to the original question. Our approach is shown to find many correct answers, even under these circumstances.

Table 2: Akron Sample Candidate Answers.

th a population of **217,074** (2000 census).<div/
n, OH. Population: **214,556**. Compare Akron to T
as a population of **200,000**. Through the 1900s,
any) was producing **200,000** "Diamond" brand lam
ad a population of **199,110**. The Akron Metropol

4.3 City Population Results

Using the Baseline algorithm described in Section 3.3, we get the results for Abilene as shown in Table 3.

$$Precision = (0.668/1 + 0.175/4) / 2 = 0.356$$

Using this baseline technique, we achieve an overall mean precision of 0.233 for all 149 cities.

Table 3: Top Baseline Results for Abilene.

Rank	Candidate Answer	Weight	City Rightiness	MSA Rightiness
1	117,063	2	0.668	0.175
2	11,928,500	2	0.001	0.002
3	1,000,000	2	0.018	0.026
4	116,966	2	0.656	0.175

Applying the measures defined in Section 3, we derive Table 4. Only the first three rows are shown because the remaining values have smaller Rightiness values, especially when divided by rank.

$$Precision = (0.431/1 + 0.169/2) / 2 = 0.258$$

For the entire sample of 149 cities, the mean precision score was 0.297. We expect overall performance to improve as the number of samples increases, because we expect that the "noise" values will have more random distribution than the values that match the targets.

[3] wikipedia: List_of_Metropolitan_Statistical_Areas
[4] wikipedia: List_of_United_States_cities_by_population
[5] wikipedia: State_populations
[6] wikipedia: List_of_U.S._states_and_territories_by_population

Table 4: Top Single Results for Abilene.

Rank	Candidate Answer	Support	City Rightiness	MSA Rightiness
1	112,205	4.600	0.431	0.162
2	114,757	4.572	0.510	0.169
3	109,687	4.468	0.380	0.155
4	116,966	4.245	0.656	0.175

Sorting all the pairs of values found in two or more domains yields Table 5.

We score the Paired Result for Abilene as:

```
Precision = (0.356 + 0.148) / 2 = 0.252.
```

Using this technique, we achieve an overall mean precision of 0.291 for all 149 cities.

Table 5: Top Paired Results for Abilene.

Rank	Candidate Answer Pair	Paired Support	City R'ness	MSA R'ness
1	106,654 - 108,157	4.273	0.356	0.148
2	975,619 - 1,005,656	3.394	0.018	0.027
3	108,157 - 116,966	2.644	0.656	0.152

Combining the two techniques achieves the best results. For each pair as reported above, the Single Results are added to the top Paired Results, yielding the results in Table 6. For example, the support value for 116,966 is 4.245 (from Table 4), plus 2.644 (from Table 5) for a total of 6.889.

```
Precision = (0.656 + 0.026) / 2 = 0.341
```

Based on this process, we arrive at a mean precision of 0.319.

Table 6: Combined Results for Abilene.

#	Candidate Answer	Support	City Rightiness	MSA Rightiness
1	116,966	6.889	0.656	0.175
2	1,005,656	5.021	0.018	0.026

4.4 City Population Summary

We have explored three different techniques for finding multiple answers when it is known that there may be more than one correct answer. For city populations, all three techniques show a significant improvement over the baseline case, as shown in Table 7.

Table 7: City Population Results Summary.

	Baseline Results	Single Results	Paired Results	Combined Results
Precision	0.233	0.297	0.291	0.319
Improvement	-	31%	29%	41%

4.5 Country Area Results

This question is similar to the city population question in that it will often have at least two distinct answers.

q9: *What is the area of New Zealand?*

There are 219 countries being studied. Each has two primary measurements for country areas, land-only, and total-including-water. For some countries, such as Kuwait or Luxembourg, these measures will be virtually identical. But for those with interior lakes and rivers, the measurements will differ. The interior of the United States, for example, has almost 3.8 million square miles, of which about 6% is water[7].

[7] wikipedia: United_States#Geography

Validating our approach with country areas, we find that all three techniques out-performed the baseline case by a significant margin, as shown in Table 8.

Table 8: Country Area Results Summary.

	Baseline Results	Single Results	Paired Results	Combined Results
Precision	0.297	0.355	0.398	0.347
Improvement	-	20%	34%	17%

5. NAMED ENTITY METHODOLOGY

The methodology from the previous sections has been adapted to handle named entities instead of numbers. We analyze questions with multiple target answers, specifically movie star names in Section 6 and basketball player names in Section 6.5.

5.1 Named Entities

In both domains used for this part of the analysis, we are looking for names of people. The first step we use is to set up a series of regular expressions to help locate potential names within the search snippets returned from the Google search.

Consider the following question:

q10: *Who was the 42nd President of the United States?*

Candidate answers such as "Bill Clinton", "William Jefferson Clinton", "President Clinton", etc. all refer to the same entity. Considerable research has been done to resolve each of these names to the same person [9].

For this research, we compare candidate answers to see if they are the same person. The first heuristic is to use Levenshtein's edit distance [10] function on the whole names. If the edit distance is less than the threshold, the names are considered to represent the same person. The threshold is two if the shorter name is at least five characters long, otherwise it is one. If the edit distance was larger than the threshold, we use Levenshtein's edit distance on a word-level rather than character-level.

Using these two heuristics, we were able to collect a sufficient number of names for analysis, even though it is imperfect in both directions. It will occasionally miss two names that represent the same person, and will also occasionally match two different people if their names are quite similar.

5.2 Two Levels of Searches

For movie stars, the results were heavily influenced by movie directors; similarly, the basketball star results were heavily influenced by coach names. Rather than excluding these names through some offline process, we chose to use the identical approach to find the directors (coaches) first, then search for stars (players) second, and exclude the former results from the latter.

This gave us additional validation of our approach, and it also significantly improved our precision. We report both sets of results for both sample datasets. All use identical processing, and all were developed based on a 10% sample.

5.3 Answer Selection

We use a frequency counting technique for rank ordering the candidate answers. Names that match according to the results in Section 5.1 are given a scaled support value of 1.0 regardless of the edit distance value. This is because names occur with so many variations for the same person.

Once the candidate answers are rank ordered, a selection process is followed that depends on the amount of support for the top candidate answer. These numbers are empirically derived and are meant to find the top answers that are separated from the rest of the answers. Table 9 shows how the threshold is calculated.

Table 9: Support Threshold Calculation.

Highest Support	Offset	Percentage	Of the Amount Over
≤ 6	0	80%	-
> 6 and ≤ 20	4.8	50%	6
Otherwise	11.8	30%	20

For example, if the highest support value is 5, then all candidate answers with a support value of at least 4 (80% of 5) are chosen as answers. If the highest support value is 50, then all candidate answers with a support of at least 11.8 + 30% * (50-20), which is 20.8, are chosen.

This ad hoc approach was chosen based on empirical analysis. Given an ordered set of candidate responses, we wanted to choose all the candidate answers within the "top" group, with the hope of finding some separation in support values between the "top" group and the other candidate answers.

5.4 Measuring Correctness

For directors (coaches), we choose all the candidate answers that are above the threshold and declare them as consensus answers. Then we check to see if they are in the list of target answers (normally just one) using the technique in Section 5.1. From here, we simply count the number of right and wrong consensus answers.

For movie stars (basketball players), we do the same, except we exclude all candidate answers that were chosen as directors (coaches). We exclude based on our process, not the target answers.

5.5 Paired Support

As described in Section 3.5, we look for the co-occurrence of pairs of candidate answers across two different domains. If a pair of names are found in two different domains, both names are used as consensus answers, regardless of their support values.

We are only looking at the snippets as returned by Google, not the entire document. These snippets are just a few lines long and are not likely to include many different candidate answers.

6. MOVIE STAR RESULTS

6.1 Movie Star Target Answers

We used a list of 500 "must see" movies from the Internet Movie Database [11]. In our dataset there are 414 cases with only one director, 71 have two directors, and in 15 cases there are three directors Each movie is annotated with four movie stars as well as the director(s). No movies were filtered from the list, even though many have unusual names ("Network", "M", "Once", "54" and others). Table 10 shows the first two movies from the dataset.

Our process does not require all four movie stars to match. We require at least one movie star and no more than four. We do not ascribe any significance to the order of the four movie stars in the original source. Likewise, we know that our process will not find movie stars if they are also the director (Mel Gibson in Braveheart, for example).

Table 10: Sample Movie Directors and Stars.

Movie	Director(s)	Stars
Ben-Hur	William Wyler	Charlton Heston, Jack Hawkins, Stephen Boyd, Haya Harareet
Gone with the Wind	Victor Fleming	Clark Gable, Vivien Leigh, Thomas Mitchell, Barbara O'Neil

6.2 Movie Director Results

A sample search string we sent to Google is:

```
director movie "Gone with the Wind" num=500
```

Based on the calculations expressed in Section 5, we got the top three results as shown in Table 11 for this movie.

Table 11: Director of Gone with the Wind.

Support	Candidate Answer	Result
37	Victor Fleming	correct
21	George Cukor	wrong
12	Sam Wood	(skip)

Plugging 37 into Table 9, the threshold for this question is 16.9, so we choose both Victor Fleming and George Cukor as our consensus answers. It turns out that George Cukor was actually the first director of the movie and was fired from that position[8], so this is a reasonable error. Even so, we give this question a precision score of 1 / 2 or 50%.

Across all 500 questions, 596 names were chosen as candidate answers, resulting in a mean precision of 88.73%. In 412 cases, there were only one or two directors, and we got them correct. In 67 cases, we got at least one director right, but also at least one wrong. In the remaining 21 cases, all of our guesses were wrong.

6.3 Movie Star Search Results

A sample search string we sent to Google is:

```
star movie "Gone with the Wind" num=500
```

Based on the same calculations as Section 6.2, we got the top three results as shown in Table 12 for this movie.

Table 12: Stars of Gone With The Wind.

Support	Candidate Answer	Result
10	Vivien Leigh	correct
5	Clark Gable	(skip, was correct)
3	Victor Fleming	(skip, director)

The threshold for this question is 6.8, so we choose only Vivien Leigh as our consensus answer. We give this question a precision score of 100%.

Across all 500 questions, 1,682 names were chosen as candidate answers, with a mean precision of 85.70%. In 396 cases, all the stars we chose were correct. In 58 cases, we got at least one star right, but also at least one wrong. In the remaining 46 cases, all of our guesses were wrong.

6.4 Movie Star Paired Support Results

For each movie, all pairs of names are collected for each domain. An average of 889 pairs were collected per movie (The movies M, The Game, and The Killing each had well over 15,000 pairs).

Out of 500 movies in the dataset, 32 of them had pairs of names that co-occurred in more than one domain. We find a net increase in precision from 85.13% to 87.05%. Considering only the 32 cases where at least two co-occurring pairs were found, 23 additional correct consensus answers were found (for 17 movies) and 4 incorrect consensus answers were found (for 3 movies). Table 13 gives justification for the wrong answers produced by paired support.

[8] wikipedia: George_Cukor

Table 13: Paired Support Impact for Movies.

Movie	Justification	Impact
JFK	Tom Hanks is the director of a second JFK movie. Lee Harvey Oswald was the assassin.	2 wrong
The Fighter	The movie is about Mickey Ward	1 wrong
V for Vendetta	Alan Moore is the author of the book	1 wrong

Note that all three errors are still reasonable. One is the director for another movie, one is the subject of the movie, and the other is the author of the book that the movie was based on.

In summary, our approach only got co-occurring pairs of candidate answers in 32 out of 500 cases. In 17 cases (53%) we had improved results, but in 3 cases (9%) our results got worse (although justifiable, as in Table 13).

6.5 Basketball Player Results

To validate the results of Section 5, we used the same process for another domain, college basketball players and coaches. All teams in our dataset have one coach, with a grand total of 4,442 players. Across all 330 questions, 429 coaches and 936 players were selected as consensus answers. The mean precision was 47.86%.

Only two questions had pairs of names that co-occurred in more than one domain. Table 14 shows the net influence of the paired support algorithm on these questions. Because so few co-occurring pairs were detected, there was very little opportunity to significantly improve the results. Even so, it did have a slight positive impact in both cases.

Table 14. Paired Support Impact for Basketball Players

Team	Player Names	Impact
Utah Utes	Jason Washburn and David Foster	2 right
Virginia Cavaliers	Darion Atkins and Evan Nolte	2 right

7. CONCLUSIONS

In this paper, we have developed several techniques for searching the web when it is known that there may be more than one correct answer. Our experiments show that our techniques consistently improve search precision.

The improvement for country areas is not quite as large as city populations because many countries have just one value for both land area and total area. For movie stars and basketball players, our search for co-occurring pairs of candidate answers did not find as many pairs as we anticipated. Most likely, this is due to the fact that we are searching just the two or three line snippet returned by Google. It is unlikely that many names will appear in such a small snippet.

Overall, detecting pairs of values that co-occur on different website domains was shown to improve search performance, when the question can have multiple correct answers.

8. REFERENCES

[1] X. Yin, W. Tan and C. Liu, "FACTO: A Fact Lookup Engine Based on Web Tables," in *World Wide Web Conference (WWW)*, Hyderabad, India, 2011.

[2] M. Wu and A. Marian, "Corroborating Answers from Multiple Web Sources," in *The 10th International Workshop on Web and Databases*, Beijing, China, 2007.

[3] J. G. Kooij, "Ambiguity in Natural Language: an Investigation of Certain Problems in its Linguistic Description," North Holland Linguistic Series, Amsterdam, 1971.

[4] R. Kass and T. Finin, "Modeling the User in Natural Language Systems," *Computational Linguistics,* vol. 14, no. 3, pp. 5-22, 1988.

[5] S. O'Hara and T. Bylander, "Numeric Query Answering on the Web," *International Journal on Semantic Web and Information Systems,* pp. 1-17, January-March 2011.

[6] S. O'Hara and T. Bylander, "Predicting Website Correctness from Consensus Analysis," in *Research in Applied Computation Symposium*, San Antonio, TX, 2012.

[7] D. Radev, W. Fan, H. Qi, H. We and A. Grewal, "Probabilistic question answering on the Web," *Journal of the American Society for Information Science and Technology,* vol. 566, pp. 571-583, 2005.

[8] D. Cohn, "State Population Estimates and Census 2010 Counts: Did They Match?" Pew Research, 12 January 2011. [Online]. Available: http://www.pewsocialtrends.org/2011/01/12/state-population-estimates-and-census-2010-counts-did-they-match/. [Accessed 28 August 2013].

[9] M. Barcala, J. Vilares, M. A. Alonso, J. Grana and M. Vilares, "Tokenization and Proper Noun Recognition for Information Retrieval," Departamento de Computacion, Universidade da Coruna, La Coruna, Spain, 2002.

[10] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Cybernetics and Control Theory,* pp. 845-848, 1965.

[11] Slavlov-Daniel, "Top 500 Must-See Movies," 13 February 2011. [Online]. Available: http://www.imdb.com/list/nA5Lm3rUIxs/. [Accessed 12 May 2013].

[12] "United States - 2013, CIA World Fact Book," 2013. [Online]. Available: http://www.theodora.com/wfbcurrent/united_states/index.html. [Accessed 28 August 2013].