Overview of the Triple Scoring Task at the WSDM Cup 2017

Hannah Bast
University of Freiburg
bast@cs.uni-freiburg

Björn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg

Elmar Haussmann
University of Freiburg
haussmann@cs.uni-freiburg

ABSTRACT

This paper provides an overview of the triple scoring task at the WSDM Cup 2017, including a description of the task and the dataset, an overview of the participating teams and their results, and a brief account of the methods employed. In a nutshell, the task was to compute relevance scores for knowledge-base triples from relations, where such scores make sense. Due to the way the ground truth was constructed, scores were required to be integers from the range 0..7. For example, reasonable scores for the triples *Tim Burton profession Director* and *Tim Burton profession Actor* would be 7 and 2, respectively, because Tim Burton is well-known as a director, but he acted only in a few lesser known movies.

The triple scoring task attracted considerable interest, with 52 initial registrations and 21 teams who submitted a valid run before the deadline. The winning team achieved an accuracy of 87%, that is, for that fraction of the triples from the test set (which was revealed only after the deadline) the difference to the score from the ground truth was at most 2. The best result for the average difference from the test set scores was 1.50.

1. INTRODUCTION

Knowledge bases allow queries that express the search intent precisely. For example, we can easily formulate a query that gives us precisely a list of all *American actors* in a knowledge base. Note the fundamental difference to full-text search, where keyword queries are only approximations of the actual search intent, and thus result lists are typically a mix of relevant and irrelevant hits.

But even for result sets containing only relevant items, a ranking of the contained items is often desirable. One reason is similar as in full-text search: when the result set is very large, we cannot look at all items and thus want the most "interesting" items first. But even for small result sets, it is useful to show the inherent order of the items in case there is one. We give two examples. The numbers refer to a sanitized dump of Freebase from June 29, 2014; see [2].

Example 1 (American actors): Consider the query that returns all entities that have *Actor* as their profession and *American* as their nationality. On the mentioned version of Freebase, this query has 64,757 matches. A straightforward ranking would be by popularity, as measured, e.g., by counting the number of occurrences of each entity in a reference text corpus. Doing that, the top-5 results for our query look as follows (the first result is G. W. Bush):

George Bush,Hillary Clinton,Tim Burton,Lady Gaga,Johnny Depp

All five of these are indeed listed as actors in Freebase. This is correct in the sense that each of them appeared in a number of movies, and be it only in documentary movies as themselves or in short cameo roles. However, Bush and Clinton are known as politicians, Burton is known as a film director, and Lady Gaga as a musician.

Only Johnny Depp, number five in the list above, is primarily an actor. He should be ranked before the other four.

Example 2 (professions of a single person): Consider all professions by Arnold Schwarzenegger. Freebase lists 10 entries:

Actor, *Athlete*, *Bodybuilder*, *Businessperson*, *Entrepreneur*, *Film Producer*, *Investor*, *Politician*, *Television Director*, *Writer*

Again, all of them are correct in a sense. For this query, ranking by "popularity" (of the professions) makes even less sense than for the query from Example 1. Rather, we would like to have the "main" professions of that particular person at the top. For Arnold Schwarzenegger that would be: *Actor*, *Politician*, *Bodybuilder*. Note how we have an ill-defined task here: it is debatable whether Arnold Schwarzenegger is more of an actor or more of a politician. But he is certainly more of an actor than a writer.

1.1 Task Definition

The task is to compute relevance scores for triples from type-like relations. The following definition is adapted from [3]:

Definition: Given a list of triples from two type-like relations (profession and nationality), for each triple compute an integer score from 0..7 that measures the degree to which the subject belongs to the respective type (expressed by the predicate and object).

Here are four example scores, some of which are related to the example queries above. The numbers in parentheses are the respective scores normalized to the more intuitive range $[0, 1]$ (that is, divided by 7) and rounded to two digits after the dot.

Tim Burton	*profession*	*Director*	7	(1.00)
Tim Burton	*profession*	*Actor*	3	(0.43)
Johnny Depp	*profession*	*Actor*	7	(1.00)
Roger Federer	*nationality*	*Swiss*	7	(1.00)
Roger Federer	*nationality*	*South African*	1	(0.14)

An alternative way of expressing this notion of degree is: How "surprised" would we be to see, for example, Roger Federer in a list of South Africans. This information is correct, yet most people would be very surprised, hence the low relevance score above. The "surprised" formulation is also used in the crowdsourcing task which we designed to acquire human judgments for the ground truth used in our evaluation.

1.2 Datasets and Evaluation

Participants were provided a knowledge base in the form of 818,023 triples from two Freebase relations: *profession* and *nationality*. Overall, these triples contain 376,214 different subjects, 200 different professions, and 100 different nationalities.

We constructed a ground truth for 1,387 of these triples (1,028 profession, 359 nationality). For each triple, we obtained 7 binary relevance judgments from a carefully implemented and controlled

crowdsourcing task, as described in [3]. This gives a total of 9,709 relevance judgments. For each triple, the sum of the binary relevance judgments yields the score, which can be any integer in the range 0..7.

About half of this ground truth (677 triples) was made available to the participants as training data. This was useful for understanding the task and the notion of "degree" in the definition above. However, the learning task was still inherently unsupervised, because the training data covers only a subset of all professions and nationalities. Participants were allowed to use arbitrary external data for unsupervised learning. For convenience, we provided 33,159,353 sentences from Wikipedia for most (all but 407) subjects from the knowledge base. For each subject from the ground truth, there was at least one sentence (and usually many more) with that subject annotated. Not surprisingly, all participating teams made use of this data.

The submissions were evaluated on a test set consisting of 710 triples. These triples were known to be a subset of the 818,023 triples from the provided knowledge base. However, it was (of course) not known before the submission deadline, which subset this was.

Three quality measures were applied to measure the quality of participating systems with respect to our ground truth:

Accuracy (ACC): the percentage of triples for which the score (an integer from the range 0..7) differs by at most 2 (in either direction) from the score in the ground truth.

Average score difference (ASD): the average (over all triples in the ground truth) of the absolute difference of the score computed by the participating system and the score from the ground truth.

Kendall's Tau (TAU): a rank-based measure which compares the ranking of all the professions (or nationalities) of a person with the ranking computed from the ground truth scores. The handling of items with equal score is described in [3, Section 5.1] and under the link at the end of this section.

The winners of the competition (there was a money prize for the top 3) were determined according to the ranking with respect to the ACC measure. If a team submitted more than one valid run, the last valid run submitted before the deadline was considered.

It should be noted that the ACC measure can only increase (and never decrease) when all scores 0 and 1 are rounded up to 2, and all scores 6 and 7 are rounded down to 5. We call this the *2-5-trick* in the following. For winning the competition, it was advantageous to apply this trick, but (as we will see in Section 3) the other measures suffer when applying this transformation.

Some teams applied this trick, but most teams did not. In Section 3, we therefore show three result tables: the results for the official submissions (with some teams using the trick, and most not), the results where the trick is used by no one, and the results where the trick is used by everyone. As discussed in Section 4, in retrospect, it would have been better to take ASD as the measure for determining the winners, since it is strongly related to ACC but does not benefit from the 2-5-trick. However, the winning team is the same with respect to both ACC and ASD, and so is the runner-up.

The setup, datasets, rules, and measures, are also described in detail on the website of the triple scoring task: http://www.wsdm-cup-2017.org/triple-scoring.html.

1.3 Related publications

The WSDM Cup 2017 offered two tasks: vandalism detection and triple scoring. The two tasks were completely independent, and only loosely related in the sense that both address a fundamental challenge when working with a very large knowledge base.

In [11], a brief overview of both tasks was provided. The deadline for that overview paper was before the submission deadline for both tasks, hence it contains no information on the participating teams and the main results.

The triple scoring problem was introduced in [3]. The paper describes how the ground truth is obtained (via crowd-sourcing), and it presents several approaches (old and new) for solving the problem as well as an extensive evaluation. A survey of basic and advanced techniques for entity search, and more generally semantic search, is provided in [4]. Both papers were known to the participants.

The participating teams developed several new ideas for solving the problem. 13 teams submitted a notebook paper describing their approach, see Table 1.

2. PARTICIPATING TEAMS

Overall, 52 teams registered on the website, of which 33 also registered on TIRA (the platform we used for submitting runs in a reproducible fashion). Eventually, 21 teams made a valid submission before the deadline. The names and affiliations of these 21 teams are listed in Table 1. The team names were randomly assigned by the task organizers (us) from a list of healthy vegetables.

Team	Affiliation	Country
Bokchoy [7]	Chinese Academy of Sciences	CN
Bologi	University of Illinois (UIUC)	US
Cabbage [5]	Negev + Tel Aviv University	IL
Catsear [13]	University of Leipzig	DE
Cauliflower	Yahoo! Japan	JP
Celosia [9]	IIIT Hyderabad + Microsoft	IN
Chaya	KAIST	KR
Chickweed	Austral University	AR
Chicory [8]	Radboud University + Spinque	NL
Cress [10]	NTNU Trondheim + U Stavanger	NO
Endive		IR
Fiddlehead [14]	Fuji Xerox	JP
Gailan [1]	Trinity College	IE
Goosefoot [17]	Sofia University + QCRI	BG+QA
Kale	University of Mannheim	DE
Lettuce [16]	Studio Ousia + Nara Institute	JP
Pigweed [12]	Yahoo! Japan	JP
Radicchio [6]	University of Illinois (UIUC)	US
Rapini		US
Samphire [15]	Indiana University Bloomington	US
Yarrow	University of Illinois (UIUC)	US

Table 1: The 21 teams (in alphabetical order) who submitted a valid run, with their affiliations and publication if available.

3. MAIN RESULTS

Table 2 shows the top 10-results of the official submissions with respect to each of the three quality measures explained in Section 1.2. As explained in Section 1.2, if a team submitted several runs, the last run that was submitted before the deadline was taken.

#	Team	ACC	#	Team	ASD	#	Team	TAU
1.	Bokchoy*	0.868	1.	Cress	1.613	1.	Goosefoot	0.314
2.	Lettuce*	0.823	2.	Bokchoy*	1.630	2.	Cress	0.321
3.	Radicchio	0.797	3.	Radicchio	1.692	3.	Bokchoy*	0.327
4.	Catsear	0.796	4.	Fiddlehead	1.704	4.	Chaya	0.337
5.	Samphire*	0.780	5.	Cabbage	1.735	5.	Chicory	0.353
6.	Cress	0.779	6.	Lettuce*	1.762	6.	Cabbage	0.354
7.	Chickweed	0.772	7.	Goosefoot	1.776	7.	Lettuce*	0.362
8.	Cauliflower*	0.752	8.	Chaya	1.811	8.	Kale	0.363
9.	Goosefoot	0.747	9.	Gailan	1.837	9.	Gailan	0.370
10.	Cabbage	0.737	10.	Kale	1.855	10.	Chickweed	0.392
14.	Baseline	0.721	21.	Baseline	2.070	20.	Baseline	0.460

Table 2: Top-10 results of the official submissions and the baseline. The teams marked * rounded all scores to the range 2..5 (the 2-5-trick described in Section 1.2) for their official submission.

#	Team	ACC	#	Team	ASD	#	Team	TAU
1.	Bokchoy	0.818	1.	Bokchoy↑	1.501	1.	Lettuce↑	0.294
2.	Radicchio↑	0.797	2.	Lettuce↑	1.594	2.	Goosefoot↓	0.314
3.	Catsear↑	0.796	3.	Cress↓	1.613	3.	Bokchoy	0.316
4.	Cress↑	0.779	4.	Radicchio↓	1.692	4.	Cress↓	0.321
5.	Lettuce↓	0.772	5.	Fiddlehead↓	1.704	5.	Chaya↓	0.337
6.	Chickweed↑	0.772	6.	Cabbage↓	1.735	6.	Cabbage	0.354
7.	Goosefoot↑	0.746	7.	Goosefoot	1.776	7.	Kale↑	0.363
8.	Cabbage↑	0.737	8.	Chaya	1.811	8.	Gailan↑	0.370
9.	Pigweed↑	0.737	9.	Gailan	1.837	9.	Chickweed↑	0.392
10.	Fiddlehead↑	0.728	10.	Kale	1.855	10.	Fiddlehead↑	0.395
11.	Baseline↑	0.721	20.	Baseline↑	2.070	20.	Baseline	0.460

Table 3: Top-10 results of the submissions with scores in the full 0..7 range, as well as the (unchanged) baseline. The arrows indicate if the ranking of a team improved or worsened compared to the respective column in Table 2

#	Team	ACC	#	Team	ASD	#	Team	TAU
1.	Bokchoy	0.868	1.	Bokchoy↑	1.630	1.	Bokchoy↑	0.327
2.	Lettuce	0.823	2.	Cabbage↑	1.735	2.	Chaya↑	0.332
3.	Radicchio	0.814	3.	Radicchio	1.754	3.	Cress↓	0.335
4.	Cabbage↑	0.813	4.	Goosefoot↑	1.755	4.	Goosefoot↓	0.343
5.	Gossefoot↑	0.804	5.	Lettuce↑	1.762	5.	Cabbage↑	0.359
6.	Cress	0.797	6.	Cress↓	1.768	6.	Lettuce↑	0.362
7.	Catsear↓	0.796	7.	Fiddlehead↓	1.820	7.	Chicory↓	0.368
8.	Chaya↑	0.786	8.	Bologi↑	1.823	8.	Bologi↑	0.385
9.	Filddehead↑	0.782	9.	Chaya↓	1.827	9.	Chickweed↑	0.386
10.	Chickweed↓	0.782	10.	Catsear↑	1.859	10.	Kale↓	0.389
18.	Baseline↓	0.721	21.	Baseline	2.070	20.	Baseline	0.460

Table 4: Top-10 results of the submissions with scores rounded to the range 2..5, as well as the (unchanged) baseline. The arrows indicate if the ranking of a team improved or worsened compared to the respective column in Table 2.

Included are the results for a very simple *baseline* method, which simply assigns score 5 to each triple. This is a reasonable baseline, because this is the one fixed score that gives the best result on the training set with respect to ACC. The reason is that the majority of triples are relevant and that with respect to ACC the score 5 is the best choice for a relevant triple, since it counts as accurate when the score in the ground truth is 3, 4, 5, 6, or 7.

Since the three official winners of the competition were selected according to ACC (first column of Table 2), several teams exploited the 2-5-trick explained in Section 1.2: computing only scores from the range 2..5. As explained above, this can (and usually does) harm ASD and TAU, but can only improve ACC and never make it worse. The teams who exploited this trick are marked with a * in Table 2.

As some teams exploited this trick to improve their chances of winning the prize money (which was a completely reasonable thing to do) yet most teams did not, the results from these two groups are somewhat hard to compare. We therefore also provide an evaluation of two variants of the submitted runs.

Table 3 shows the results, when all teams use the full score range. For this table, the teams marked * in Table 2 were individually asked (outside of the competition) to run a variant of their method that uses the full score range (thus likely making their ACC scores worse, but likely improving their ASD and TAU scores).

Table 4 shows the results, when all the scores from the official submissions were truncated to the range 2..5. This transformation was trivial to apply (and did not change anything for the teams marked * in Table 2).

4. DISCUSSION

With respect to ACC, almost all teams that did not use the 2-5-trick benefit quite significantly from the truncation to the range 2..5. For the top-10 teams from the first column (ACC) of Table 2, the improvement ranges from 0 (Catsear) to 0.076 (Cabbage).

Team Bokchoy, which is the winner with respect to ACC, is also in the top-3 with respect to ASD and TAU, despite using the 2-5-trick. With scores in the full range for all teams, team Bokchoy also wins with respect to ASD, and with scores in the range 2..5 for all teams, team Bokchoy also wins with respect to TAU. This demonstrates the particular strength of their approach, which is briefly described in Section 5 below.

Performance with respect to ACC and ASD is strongly related for all teams. This is intuitive, because both measures require a good estimate of the score from the ground truth.

Performance with respect to TAU is only weakly related to performance with respect to ACC or ASD. For example, none of the teams on places 3, 4, and 5 with respect to ACC are in the top-10 with respect to TAU. Indeed, ranking is a relatively simpler task than estimating the score. For example, when a subject has only two objects, finding the right ranking is a binary classification problem, whereas estimating the score has more degrees of freedom.

For future evaluations of this task, ASD should be preferred to ACC, because ASD does not benefit from the 2-5-trick, but is otherwise strongly related to ACC. TAU is a measure of independent interest. If only one measure should be singled out, ASD is the measure of choice, because minimizing ASD is harder than minimizing TAU. However, it should be noted that some applications may only require a ranking and not the exact scores, in which case TAU is a perfectly reasonable measure.

The simple baseline beats one third of the teams with respect to ACC. This is simply due to the fact, that many teams did not use the 2-5-trick (and thus did not optimize strictly for ACC). With respect to ASD and TAU, the baseline ranks poorly, as it should.

5. METHODS EMPLOYED BY THE PARTICIPANTS

All 21 teams who submitted a valid run made use of the Wikipedia sentences provided by the organizers; see Section 1.2. Many (but not all) teams used variants of the ideas presented in [3]. For details, we refer the interested reader to the various notebook papers referenced in Table 1.

We briefly summarize the approach of team Bokchoy [7], who did not only win the competition, but were among the top-3 with respect to all three quality measures (ACC, ASD, TAU) and in all three variants of our evaluation (Tables 2, 3, and 4). Their approach has two main components.

The first component is an ensemble learner using four independent approaches for computing triple scores. Three of these approaches were taken from the original paper [3], all using the Wikipedia sentences. The fourth approach makes use of Freebase paths between the subject and object of the triple to be scored. These paths are used as features for a binary classifier that decides whether the triple should get a high score or a low score. For example, the path *Johnny Depp born-in Kentucky located-in USA* is a strong indication that the triple *Johnny Depp nationality USA* should get a high score.

The second component can promote or demote scores based on so-called trigger words for a type (which in this task was either a nationality or a profession). These are words that are semantically very strongly related to the type. For example, trigger words for the German nationality are *German* and *Germany*. Trigger words were compiled from a variety of sources, including a list of adjectival forms, synonyms and hyponyms from WordNet, and some manually compiled words. The trigger words were then employed as follows to modify the score s output by the ensemble learner, using the first paragraph of the Wikipedia article of the subject of the triple in question. If *the first sentence* of that paragraph contains at least one of the trigger words, promote the score to $\max\{5, s\}$. If *the whole paragraph* contains none of the trigger words, demote the score to $\min\{2, s\}$.

The intuition behind this second component is that a mention of a trigger word in the first sentence of a Wikipedia article is a strong signal for a high relevance of the respective information. Similarly, if no trigger word is mentioned in the whole first paragraph (which in case of a person entity is usually a synopsis of the main aspects of that person's life), this is a strong signal for a low relevance of the respective information. This simple post-processing of the scores improves the already good quality of the ensemble scorer significantly.

6. CONCLUSION

We have presented and discussed the results of the triple scoring task of the WSDM Cup 2017. Triple scoring is a basic ingredient in entity ranking, that is, for searches that return a list of entities. The task has attracted considerable interest from all over world, with 52 initial registrations and 21 teams submitting a valid run before the deadline. A summary of the task and all the relevant data is freely available from the task's website: http://www.wsdm-cup-2017.org/triple-scoring.html[1]. We hope that the task and this overview and the free availability of the data spark further research on this important and interesting problem.

---

[1] Should this URL ever cease to function, search for the keywords *triple scoring wsdm cup 2017* and you should find another page with the same content.

References

[1] E. Ali, A. Caputo, and S. Lawless. Triple scoring using paragraph vector - the Gailan triple scorer. In *WSDM Cup*, 2017.

[2] H. Bast, F. Bäurle, B. Buchhold, and E. Haußmann. Easy access to the freebase dataset. In *WWW*, pages 95–98, 2014. URL http://doi.acm.org/10.1145/2567948.2577016.

[3] H. Bast, B. Buchhold, and E. Haussmann. Relevance scores for triples from type-like relations. In *SIGIR*, pages 243–252, 2015. URL http://doi.acm.org/10.1145/2766462.2767734.

[4] H. Bast, B. Buchhold, and E. Haussmann. Semantic search on text and knowledge bases. *Foundations and Trends in Information Retrieval*, 10(2-3):119–271, 2016. URL https://doi.org/10.1561/1500000032.

[5] Y. Brumer, B. Shapira, L. Rokach, and O. Barkan. Predicting relevance scores for triples from type-like relations using neural embedding - the Cabbage triple scorer. In *WSDM Cup*, 2017.

[6] L.-W. Chen, B. Mangipudi, J. Bandlamudi, R. Sehgal, Y. Hao, M. Jiang, and H. Gui. Integrating knowledge from latent and explicit features for triple scoring - the Radicchio triple scorer. In *WSDM Cup*, 2017.

[7] B. Ding, Q. Wang, and B. Wang. Leveraging text and knowledge bases for triple scoring: An ensemble approach - the Bokchoy triple scorer. In *WSDM Cup*, 2017.

[8] F. Dorssers, A. P. de Vries, W. Alink, and R. Cornacchia. Ranking triples using entity links in a large web crawl - the Chicory triple scorer. In *WSDM Cup*, 2017.

[9] N. Fatma, M. K. Chinnakotla, and M. Shrivastava. Relevance scoring of triples using ordinal logistic classification - the Celosia triple scorer. In *WSDM Cup*, 2017.

[10] F. Hasibi, D. Garigliotti, S. Zhang, and K. Balog. Supervised ranking of triples for type-like relations - the Cress triple scorer. In *WSDM Cup*, 2017.

[11] S. Heindorf, M. Potthast, H. Bast, B. Buchhold, and E. Haussmann. WSDM cup 2017: Vandalism detection and triple scoring. In *WSDM*, pages 827–828, 2017. URL http://dl.acm.org/citation.cfm?id=3022762.

[12] V. Kanojia, R. Togashi, and H. Maeda. Relevance score of triplets using knowledge graph embedding - the Pigweed triple scorer. In *WSDM Cup*, 2017.

[13] E. Marx, T. Soru, and A. Valdestilhas. Triple scoring using a hybrid fact validation approach - the Catsear triple scorer. In *WSDM Cup*, 2017.

[14] M. Sato. Predicting triple scoring with crowdsourcing-specific features - the Fiddlehead triple scorer. In *WSDM Cup*, 2017.

[15] P. Shiralkar, M. Avram, G. L. Ciampaglia, F. Menczer, and A. Flammini. Relsifter: Scoring triples from type-like relations - the Samphire triple scorer. In *WSDM Cup*, 2017.

[16] I. Yamada, M. Sato, and H. Shindo. Ensemble of neural classifiers for scoring knowledge base triples - the Lettuce triple scorer. In *WSDM Cup*, 2017.

[17] V. Zmiycharov, D. Alexandrov, P. Nakok, I. Koychev, and Y. Kiprov. Finding people's professions and nationalities using distant supervision - the Goosefoot triple scorer. In *WSDM Cup*, 2017.