Intelligent Information Management
Targeted Competition Framework
ICT-2011.4.4(d)

Project **FP7-318652 / BioASQ**

Deliverable **excerpt from D4.1**

Distribution **Public**

http://www.bioasq.org

Evaluation Measures for Task B

Prodromos Malakasiotis, Ioannis Pavlopoulos, Ion Androutsopoulos and Anastasios Nentidis

Status: Final (Version 1.1)

November 2020

1

Task B: Biomedical Semantic Question Answering

BIOASQ Task B takes place in two phases:

Phase A (annotate questions, retrieve relevant articles, snippets, triples): In this phase, participants are provided with biomedical questions written in English and are asked to: (i) semantically annotate the questions with concepts from a set of designated terminologies and ontologies; and (ii) retrieve relevant articles, text snippets and RDF triples from designated article repositories and ontologies. The designated terminologies, ontologies and article repositories are described by Tsatsaronis et al. (2013). The system responses of Phase A will be automatically compared against golden responses constructed by the BIOASQ team of biomedical experts; consult Malakasiotis et al. (2013) for information on how the golden responses will be constructed.

Phase B (find and report 'exact' and 'ideal' answers): In this phase, the questions and golden responses of Phase A (correct articles and snippets) are provided as input. The participants are asked to report 'exact answers' (e.g., named entities in the case of factoid questions) and 'ideal answers' (paragraph-sized summaries). The 'exact' and 'ideal' answers of the systems will be automatically compared against golden 'exact' and 'ideal' answers constructed by the BIOASQ team of biomedical experts; again, consult Malakasiotis et al. (2013) for information on how the golden 'exact' and 'ideal' answers will be constructed. All the 'ideal' answers of the systems will also be manually evaluated by the biomedical experts.

Phase B takes place immediately after Phase A. In both phases, the participants have very limited time to submit their responses, to make it difficult for participants to produce their responses manually.

2

Evaluation process and measures for Task b Phase A

In Phase A, the participants are provided with English questions $q_1, q_2, q_3, \ldots, q_n$. For each question q_i, each participating system is required to return:

A list of at most 10 relevant concepts $c_{i,1}, c_{i,2}, \ldots, c_{i,10}$ from the designated terminologies and ontologies. The list should be ordered by decreasing confidence, i.e., $c_{i,1}$ should be the concept that the system considers most relevant to the question q_i, $c_{i,2}$ should be the concept that the system considers to be the second most relevant etc. A single concept list will be returned per question and participant, and the list may contain concepts from multiple designated terminologies and ontologies. The returned concept list will actually contain unique concept identifiers (obtained from the terminologies and ontologies), rather than terms (words or phrases).

A list of at most 10 relevant articles (documents) $d_{i,1}, d_{i,2}, \ldots, d_{i,10}$ from the designated article repositories. Again, the list should be ordered by decreasing confidence, i.e., $d_{i,1}$ should be the article that the system considers most relevant to the question, $d_{i,2}$ should be the article that the system considers to be the second most relevant etc. A single article list will be returned per question and participant, and the list may contain articles from multiple designated repositories. The returned article list will actually contain unique article identifiers (obtained from the repositories).

A list of at most 10 relevant text snippets $s_{i,1}, s_{i,2}, \ldots, s_{i,10}$ from the returned articles. Again, the list should be ordered by decreasing confidence. A single snippet list will be returned per question and participant, and the list may contain any number (or no) snippets from any of the returned articles $d_{i,1}, d_{i,2}, d_{i,3}, \ldots, d_{i,k}$. Each snippet will be represented by the unique identifier of the article it comes from and the offsets (character positions in the article) of the snippet's beginning and end (offsets of the first and last characters).

A list of at most 10 relevant RDF triples $t_{i,1}, t_{i,2}, \ldots, t_{i,10}$ from the designated ontologies. Again, the list should be ordered by decreasing confidence. A single triple list will be returned per question and participant, and the list may contain any triples from multiple designated ontologies.

For each question q_i, the BIOASQ team of biomedical experts constructes the gold (correct) sets of articles and snippets as discussed by Malakasiotis et al. (2013). Once the responses of the participating systems have been submitted, the biomedical experts will also inspect all the concepts, articles, snippets,

and triples of each system, i.e., the 10 concepts, articles, snippets, and triples that each system is most confident about, in order to concstruct corect sets of relevant concepts and triples and to add to the corresponding golden sets of articles and snippets any correct (relevant) item that the biomedical experts had missed, but the systems managed to retrieve.

For each system, the lists of returned concepts, articles, and triples of all the questions will then be evaluated using the *mean average precision* (*MAP*) measure, defined below, which is widely used in information retrieval to evaluate ranked lists of retrieved items (see Manning et al. (2008)). We will also use the *geometric mean average precision* (*GMAP*), which places more emphasis on improvements in low performing queries (see Robertson (2006) and Sanderson (2010)). For the sake of completeness, we will also compute the *mean precision*, *mean recall*, and *mean F-measure* of each system, also defined below, but the official scores for concepts, articles, and triples in Phase A will be based on *MAP*. The list of returned snippets will also be evaluated using some versions of the same evaluation measures described above, that have been properly modified to account for potential snippet overlap as described in the following subsections. Since BioASQ9, the official scores for snippets in Phase A will be based on *mean F-measure*.

2.1 Mean precision, mean recall, mean F-measure

Given a set of golden items (e.g., articles), and a set of items returned by a system (for a particular question in our case), precision (P) and recall (R) are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{2.1}$$

$$R = \frac{TP}{TP + FN} \tag{2.2}$$

where TP (true positives) is the number of returned items that are also present in the golden set, FP (false positives) is the number of returned items that are not present in the golden set, and FN (false negatives) is the number of items of the golden set that were not returned by the system. The F_β measure is the weighted harmonic mean of P and R, defined as follows:

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R} \tag{2.3}$$

For $\beta = 1$, the same weight is assigned to both precision and recall, and the resulting measure, often called simply F-measure, is defined as follows:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{2.4}$$

Given a set of queries (in our case, questions) $q_1, \ldots, q_n$, the *mean precision*, *mean recall*, and *mean F-measure* of each system is obtained by averaging its precision, recall, and F-measure for all the queries.

In BIOASQ, we will compute the mean precision, mean recall, and mean F-measure of the concepts, articles, snippets, and triples returned by each system. In the case of snippets, a complication is that a returned snippet may overlap with one or more golden snippets, without being identical to any of them. To take this into account, in the case of snippets we modify the definitions of precision and recall. Figure 2.1 illustrates what we mean by article-offset pairs. A snippet is determined by the article it comes from and by the offsets (positions) in the article of the first and last characters of the snippet.[1] We can also

[1] Actually, snippets are also determined by the sections of the articles they come from, but for simplicity we ignore sections here. Consult the guidelines of Task B (http://participants-area.bioasq.org/) for further details.

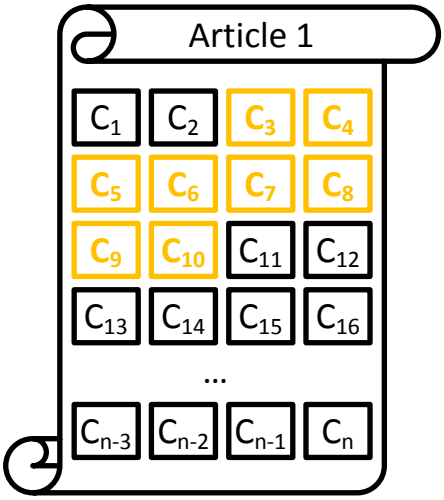

Figure 2.1: An article-offset pair example. Article 1 has n characters and a golden snippet starting at offset 3 and ending at offset 10.

think of the snippet as a set of (article, offset) pairs, one pair for each character of the snippet. In the example of Figure 2.1, Article 1 has n characters and a golden snippet starting at offset 3 and ending at offset 10. Let us call S the set of all the article-offset pairs of all the characters in the snippets returned by a system for a particular question, G the set of all the article-offset pairs of all the characters in the golden snippets of the question, and let $|s|$ denote the cardinality of a set s. The definitions of precision (P_{snip}) and recall (R_{snip}) for snippets are:

$$P_{snip} = \frac{|S \cap G|}{|S|} \tag{2.5}$$

$$R_{snip} = \frac{|S \cap G|}{|G|} \tag{2.6}$$

In effect, P_{snip} divides the size (in characters) of the total overlap between the returned and golden snippets by the total size of the returned snippets, whereas R_{snip} divides the size of the total overlap by the total size of the golden snippets. The definitions of F_β, mean precision, mean recall, and mean F-measure for snippets are the same as the corresponding definitions for concepts, articles, and triples, but they use P_{snip} and R_{snip} instead of P and R.

2.2 Mean average precision and geometric mean average precision

Precision, recall, and F-measure do not consider the order of the items returned by a system for each query. Recall that in BIOASQ we require the lists of concepts, articles, snippets, and triples that a system returns for each question to be ordered (ranked) by decreasing confidence. To take the ordering of a particular returned list (for a particular question) into account, it is common in information retrieval to compute the (non-interpolated) *average precision* (*AP*) of the list, defined as follows:

$$AP = \frac{\sum_{r=1}^{|L|} P(r) \cdot rel(r)}{|L_R|} \tag{2.7}$$

Retrieved items	Unordered retrieval measures	Ordered retrieval measures
concepts	mean precision, recall, F-measure	***MAP***, *GMAP*
articles	mean precision, recall, F-measure	***MAP***, *GMAP*
snippets	mean precision, recall, F-measure	***MAP***, *GMAP*
triples	mean precision, recall, F-measure	***MAP***, *GMAP*

Table 2.1: Evaluation measures for Phase A of Task b.

where $|L|$ is the number of items in the list, $|L_R|$ is the number of relevant items, $P(r)$ is the precision when the returned list is treated as containing only its first r items, and $rel(r)$ equals 1 if the r-th item of the list is in the golden set (i.e., if the r-th item is relevant) and 0 otherwise.[2]

Since BIOASQ 8 the following modification[3] of Equation 2.7 is used for evaluation in Phase A, to consider the limit of ten elements per question in participant submissions, as well as, the case that some questions may have less than 10 relevant elements in the golden set:

$$AP = \frac{\sum_{r=1}^{|L|} P(r) \cdot rel(r)}{min(|L_R|, 10)} \tag{2.8}$$

In BIOASQ, especially when computing the average precision of a list of *snippets*, $P(r)$ will be taken to be the snippet precision P_{snip} (as in Section 2.1) when the returned list of snippets is treated as containing only its first r snippets; and $rel(r)$ will be taken to be 1 if the r-th returned snippet has a non-zero overlap (shares at least one article-offset pair) with at least one golden snippet of the particular question.

By averaging *AP* over a set of queries (in our case, questions) $q_1, \ldots, q_n$, we obtain the *mean average precision* (*MAP*), defined as follows:

$$MAP = \frac{1}{n} \cdot \sum_{i=1}^{n} AP_i \tag{2.9}$$

where AP_i is the average precision of the list returned for query (question) q_i. In our case, each system will receive four *MAP* scores, for the lists of concepts, articles, snippets, and triples, respectively, that it returned for all the questions.

The *geometric mean average precision* (*GMAP*), defined below, is very similar to *MAP*, but it uses the geometric instead of the arithmetic mean, which places more emphasis on improvements in low performing queries, as already noted.

$$GMAP = \sqrt[n]{\prod_{i=1}^{n}(AP_i + \epsilon)} \tag{2.10}$$

An alternative way to more easily compute *GMAP* is by using the following equation:

$$GMAP = \exp(\frac{1}{n} \cdot \sum_{i=1}^{n} \ln(AP_i + \epsilon)) \tag{2.11}$$

[2] *AP* approximates the area under a recall–precision curve; consult Robertson (2006).

[3] From BIOASQ 3 and until BIOASQ 7 $|L_R|$ was set equal to 10, i.e., the maximum number of relevant items that the systems are allowed to return.

In both versions of *GMAP*, ϵ is a small number added to handle cases where $AP_i = 0$. As with *MAP*, in BioASQ each system will receive four *GMAP* scores, for the lists of concepts, articles, snippets, and triples, respectively, that it returned for all the questions. The official scores for Task b Phase A will be based on *MAP*, as already noted. Table 2.1 summarizes the evaluation measures of Phase A; the official measures are shown in bold.

3

Evaluation process and measures for Task b Phase B

In Phase B, the participants are provided with the same questions $q_1, \ldots, q_n$ as in Phase A, but this time they will also be given the golden (correct) lists of articles and snippets of each question. For each question, each participating system will have to return an 'ideal' answer, i.e., a paragraph-sized summary of relevant information. In the case of yes/no, factoid, and list questions, the systems will also have to return 'exact' answers; for summary questions, no 'exact' answers will be returned. Consult Malakasiotis et al. (2013) for a discussion of the types of questions used in BIOASQ, and the nature of 'exact' and 'ideal' answers. The participants are told the type of each question.

3.1 Evaluating 'exact' answers

We first discuss how 'exact' answers are evaluated in Phase B, by considering in turn yes/no, factoid, and list questions.

Evaluating the 'exact' answers of yes/no questions

For each yes/no question, the 'exact' answer of each participating system will have to be either 'yes' or 'no'. The response will be compared against the golden 'exact' answer (again 'yes' or 'no') that the BIOASQ team of biomedical experts will have associated with the question. For each system, we will compute the *accuracy* (*Acc*) of its responses to yes/no questions. Assuming that there are n yes/no questions, accuracy is defined as follows, where c is the number of correctly answered yes/no questions.

$$Acc = \frac{c}{n} \tag{3.1}$$

Accuracy is measured for completeness. Since BioASQ6, the official measure for the 'exact' answers of yes/no questions is based in *precision* (P), *recall* (R), and *F-measure* (F_1), as described in Section 2.1. For a set of n yes/no questions the *F-measure* (F_1) is calculated independently for "yes" and "no" answers, named F_1y and F_1n respectively. In particular, for F_1y TP is the number of questions with answer "yes" both in participant submission and golden set; FP is the number of questions with answer "yes" in the submission, but the answer in golden set is "no"; and FN is the number of questions having answer "no" in the submission, but "yes" in the golden set. Similarly, for F_1n TP

is the number of questions with answer ”no” both in submission and golden set; FP is the number of questions with answer ”no” in the submission, but ”yes” in golden set; and FN is the number of questions having answer ”yes” in the submission, but ”no” in the golden set. Finally, weighting equally system performance on ”yes” and ”no” questions, the *macro-averaged F-measure* (maF_1) is calculated as shown below which is the official measure for yes/no questions.

$$maF_1 = \frac{F_1 y + F_1 n}{2} \tag{3.2}$$

Evaluating the ‘exact’ answers of factoid questions

For each factoid question, each participating system has to return a list of up to 5 entity names, ordered by decreasing confidence. The BIOASQ team of biomedical experts will have associated with each factoid question a single golden entity name, as well as possible synonyms of that name. Once the responses of the participating systems have been submitted, the biomedical experts will also inspect the entity names returned by the participating systems for the factoid questions, in order to add synonyms they may have missed when preparing the golden answers.

We will measure the *strict accuracy* (*SAcc*) and *lenient accuracy* (*LAcc*) of each system for factoid questions. Strict accuracy counts a question as correctly answered if the golden entity name (or a synonym of that name) is the first element of the list returned by the system. By contrast, lenient accuracy counts a question as correctly answered if the golden entity name (or synonym) is included, not necessarily as the first element, in the list returned by the system. In the definitions below, n is the number of factoid questions, c_1 is the number of factoid questions that have been answered correctly when only the first element of each returned list is considered, and c_5 is the number of factoid questions that have been answered correctly in the lenient sense, when all the elements of the returned list are considered.

$$\textit{SAcc} = \frac{c_1}{n} \tag{3.3}$$

$$\textit{LAcc} = \frac{c_5}{n} \tag{3.4}$$

Strict and lenient accuracy will be measured for completeness. The official measure for the ‘exact’ answers of factoid questions will be the *mean reciprocal rank* (*MRR*), which is often used to evaluate factoid questions in question answering challenges; consult, for example, Voorhees (2001). In the definition below, for each factoid question q_i we search the returned list looking for the topmost position that contains the golden entity name (or one of its synonyms). If the topmost position is the j-th one, then $r(i) = j$; otherwise $r(i) \to +\infty$, i.e., $\frac{1}{r(i)} = 0$.

$$\textit{MRR} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{r(i)} \tag{3.5}$$

In effect, *MRR* rewards systems that manage to include the golden responses (or their synonyms) higher in the returned lists.

Evaluating the ‘exact’ answers of list questions

For each list question, each participating system has to return a list of entity names, jointly taken to constitute a single answer (e.g., the most common symptoms of a disease); for practical purposes, the maximum allowed size of each returned list may be limited (e.g., up to 100 names, each one up to 100 characters). The BIOASQ team of biomedical experts will have associated with each list question a

Question type	Participant response	Evaluation measures
yes/no	'yes' or 'no'	accuracy, **ma*F*-measure**
factoid	up to 5 entity names	strict and lenient accuracy, ***MRR***
list	a list of entity names	**mean** precision, recall, ***F*-measure**

Table 3.1: Evaluation measures for the 'exact' answers in Phase B of Task b.

golden list of entity names, also providing possible synonyms for each entity name of the golden list; consult Malakasiotis et al. (2013) for details. Again, once the responses of the participating systems have been submitted, the biomedical experts will also inspect the lists returned by the participating systems for the list questions, in order to add synonyms they may have missed.

For each list question, the list returned by the system will be compared against the golden list by computing its *precision* (P), *recall* (R), and *F-measure* (F_1), as in Section 2.1. Here TP is the number of entities that are mentioned both in the returned and the golden list; FP is the number of entities that are mentioned in the returned, but not in the golden list; and FN is the number of entities that are mentioned in the golden, but not in the returned list. If the same entity is mentioned using different synonyms in the returned and golden lists, it will be counted as having been mentioned in both lists. If an entity is mentioned multiple times, possibly using different synonyms, in the returned list, it will be counted only once.

By averaging precision, recall, and F-measure over the list questions, we will obtain the *mean average precision*, *mean average recall*, and *mean average F-measure* score of each system for list questions. The official measure for list questions will be mean F-measure. Table 3.1 summarizes the kinds of responses and the evaluation measures that will be used in Phase B.

3.2 Evaluating 'ideal' answers

For each question (yes/no, factoid, list, summary), each participating system of Phase B will also have to return a single paragraph-sized text summarizing the most relevant information of the retrieved concepts, articles, snippets, and triples of Phase A; recall that the correct concepts, articles, snippets, and triples of each question will be provided to the participants of Phase B. The returned 'ideal' answer is intended to approximate a short text that a biomedical expert would write to answer the question (e.g., including prominent supportive information), whereas the 'exact' answers are only 'yes'/'no' responses, entity names, or lists of entity names; and there are no 'exact' answers in the case of summary questions.

The BIOASQ team of biomedical experts will have associated each question with a golden 'ideal' answer. The maximum allowed length (in characters) of each 'ideal' answer to be produced by a system will be set to be equal to the maximum length of all the golden 'ideal' answers (of all the questions). The same maximum allowed length will be used for all the questions to avoid revealing the expected length of each particular 'ideal' answer.

The 'ideal' answers of the systems will be evaluated both manually (by the BIOASQ team of biomedical experts) and automatically (by comparing them to the golden 'ideal' answers). The official scores will be based on the manual evaluation; the automatic evaluation will be performed mostly to explore how well automatic evaluation measures (e.g., from multi-document text summarization) correlate with the scores of the biomedical experts.

Criterion	Explanation	Score
information recall	All the necessary information is reported.	1–5
information precision	No irrelevant information is reported.	1–5
information repetition	The answer does not repeat the same information multiple times.	1–5
readability	The answer is easily readable and fluent.	1–5

Table 3.2: Criteria for the manual evaluation of the 'ideal' answers in Phase B of Task b.

Manual evaluation of 'ideal' answers

The questions the participating systems will have to answer will be grouped according to their difficulty into three groups, i.e., "very difficult", "difficult", "easy". The difficulty of the questions will be determined by studying the performance of all the participating systems in Phase A. We expect that we will be able to classify the questions into the three groups of difficulty by examining how well the systems of Phase A managed (on average) to retrieve concepts, documents, snippets, and triples.

The BIOASQ team of biomedical experts will inspect the 'ideal' answers of the participating systems for a sample of m questions. The m questions will be from all the three groups of difficulty, with an approximately uniform distribution across the three groups. We also note that each question will have been formulated by a single biomedical expert (or by a pair of biomedical experts, for some questions) and that the biomedical expert who formulated each question should be the one to assess the 'ideal' answers of the systems for that question. Hence, care will also be taken to ensure that the sample of m questions contains approximately the same number of questions from each biomedical expert, in order to balance the workload of the biomedical experts during the manual evaluation.

Each one of the m 'ideal' answers of each system will be inspected by a biomedical expert, who will be asked to evaluate the answer in terms of *information recall* (the 'ideal' answer reports all the necessary information), *information precision* (no irrelevant information is reported), *information repetition* (the 'ideal' answer does not repeat the same information multiple times, e.g., when sentences of the 'ideal' answer that have been extracted from different articles convey the same information), and *readability* (the 'ideal' answer is easily readable and fluent). An 1–5 scale will be used in all four criteria (1 for 'very poor', 5 for 'excellent'). Table 3.2 summarizes the criteria that will be used in the manual evaluation of the 'ideal' answers in Phase B. A sample of the 'ideal' answers will be evaluated by more than one biomedical experts to measure the inter-annotator agreement.

Automatic evaluation of 'ideal' answers

The 'ideal' answers returned by the systems will also be automatically evaluated using *ROUGE*; consult Lin (2004). Roughly speaking, *ROUGE* counts the overlap between an automatically constructed summary and a set of reference (golden) summaries constructed by humans. There are several different versions of *ROUGE*. *ROUGE-N* is, defined below, uses word n-grams when computing the overlap between an automatically constructed summary S and a set $Refs$ of reference summaries:

$$\textit{ROUGE-N}(S|Refs) = \frac{\sum_{R \in Refs} \sum_{g_n \in R} C(g_n, S, R)}{\sum_{R \in Refs} \sum_{g_n \in R} C(g_n, R)} \tag{3.6}$$

In the definition above, g_n is a word n-gram, $C(g_n, S, R)$ is the number of times that g_n co-occurs in S and a reference summary R, and $C(g_n, R)$ is the number of times g_n occurs in reference R.

ROUGE-S uses skip bigrams, instead of n-grams, when computing the overlap. A skip bigram is any pair of words, maintaining the order of the two words and ignoring any intermediate words. *ROUGE-SU*

Question type	Participant response	Evaluation measures
any	paragraph-sized text	*ROUGE-2*, *ROUGE-SU4*, **manual scores**

Table 3.3: Evaluation measures for the 'ideal' answers in Phase B of Task b.

is similar to *ROUGE-S*, but it also counts unigrams (individual words) that occur both in S and $Refs$. The most widely used versions of *ROUGE* are *ROUGE-2* and *ROUGE-SU4*, which have been found to correlate well with human judgements, when multiple reference summaries are available per question; consult Lin (2004). *ROUGE-2* is *ROUGE-N* with $n = 2$; and *ROUGE-SU4* is a version of *ROUGE-SU* with the maximum distance between the words of any skip bigram limited to 4.

In BIOASQ, we will use *ROUGE-2* and *ROUGE-SU4*, with S being an 'ideal' answer constructed by a system and $Refs$ being any of the following:

- The golden 'ideal' answer of the particular question S was constructed for. Recall that there will be only one golden 'ideal' answer per question, and this may not allow *ROUGE-2* and *ROUGE-SU4* to correlate well with the scores of the manual evaluation.
- The correct snippets of Phase A for the particular question that S was constructed for.
- Pseudo-natural language renderings of the correct RDF triples for the particular question that S was constructed for.
- All the 'ideal' answers returned by the systems for the particular question that S was constructed for.
- All the 'ideal' answers returned by the systems for the particular question that S was constructed for, excluding returned 'ideal' answers that (i) were not manually evaluated by biomedical experts and (ii) were manually evaluated, but did not receive a score of at least 3 in all of the criteria of Table 3.2.
- Combinations of the above.

Table 3.3 summarizes the evaluation measures of Phase B; the official measures are shown in bold. We may also consider measures based on n-gram graphs (Giannakopoulos et al., 2008), measures that do not require reference summaries (Louis and Nenkova, 2013) or combinations of measures (Giannakopoulos and Karkaletsis, 2013) to examine if we can use them in the second BIOASQ challenge.

Bibliography

G. Giannakopoulos and V. Karkaletsis. Summary Evaluation: Together We Stand NPowER-ed. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 436–450, Karlovasi, Samos, Greece, 2013.

G. Giannakopoulos, V. Karkaletsis, G. Vouros, and P. Stamatopoulos. Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.*, 5(3):1–39, Oct. 2008. ISSN 1550-4875.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL workshop 'Text Summarization Branches Out'*, pages 74–81, Barcelona, Spain, 2004.

A. Louis and A. Nenkova. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300, 2013.

P. Malakasiotis, I. Androutsopoulos, Y. Almirantis, D. Polychronopoulos, and I. Pavlopoulos. Tutorials and Guidelines. Technical Report D3.4, BioASQ Deliverable, 2013.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*, volume 1. Cambridge University Press, 2008.

S. Robertson. On GMAP: and other transformations. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 78–83, Arlington, Virginia, 2006.

M. Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers, 2010.

G. Tsatsaronis, M. Zschunke, M. R. Alvers, and C. Plonka. Report on existing and selected datasets. Technical Report D3.2, BioASQ Deliverable, 2013.

E. Voorhees. The TREC QA Track. *Natural Language Engineering*, 7(4):361–378, 2001.