What's important in a text? An extensive evaluation of linguistic annotations for summarization

Markus Zopf, Teresa Botschen, Tobias Falke, Benjamin Heinzerling, Ana Marasović,
Todor Mihaylov, Avinesh P.V.S, Eneldo Loza Mencía, Johannes Fürnkranz and Anette Frank
Research Training Group AIPHES, TU Darmstadt / Heidelberg University, Germany
{zopf,botschen,falke,heinzerling,marasovic,mihaylov,avinesh}@aiphes.tu-darmstadt.de
{eneldo,juffi}@ke.tu-damrstadt.de frank@cl.uni-heidelberg.de

***Abstract*—Automatic text summarization aims at reducing the length of input documents while preserving the most important information. A key challenge in automatic summarization is therefore to estimate the importance of information. Most extractive summarization systems, however, usually only consider bigrams as the representation from which importance can be estimated. The potential of other text annotations such as frames or named-entities remains unexplored. In this paper, we evaluate the application potential of linguistic annotations for automatic text summarization. To this end, we extend a previously presented summarization system by replacing bigrams with a multitude of different linguistic annotation types, including n-grams, verb stems, frames, concepts, chunks, connotation frames, entity types, and discourse relation sense-types. We propose two novel evaluation methods to evaluate information importance detection capabilities. In our experiments, bigrams show the best overall performance when source document sentences have to be ranked. These results support the decision of summarization system developers to use bigrams in summarization systems. However, other annotation types perform better if the model has to distinguish between source and reference sentences.**

I. Introduction

In this paper, we investigate whether low-level linguistic annotations can improve the performance of a specific higher-level task, namely the estimation of information importance, which is a key problem for designing a good text summarization system [1]. We propose several annotation types that can be considered as 'features' upon which a summarization system can base its decision for or against the inclusion of a piece of information into a summary. We study which types of linguistic annotations prove useful to help a summarization system capture the notion of importance, and whether using such annotations as features can improve the performance of a summarization system that is specifically designed to model the notion of information importance.

An intuitive hypothesis is that for estimating information importance, linguistic features that involve abstractions over surface forms should be more apt to generalize to unseen data than surface-oriented features such as bigrams, especially when training resources are scarce or when moving to novel domains. On the other hand, linguistic annotations could also be noisy, sparse or suffer from being too fine-grained. We explore these questions by injecting knowledge from different linguistic *annotation types* (ATs), such as conceptual frames, the expression of sentiment and opinion, or discourse relations that could reflect importance.

For our experiments we build on a recently proposed summarization system [1] that is designed to learn the importance of bigrams, but replace bigrams with more abstract linguistic annotation types. This system provides an excellent test bed for our research questions: (i) the summarization task inherently relies on the notion of importance, and many approaches to summarization point out that they estimate the importance of information nuggets, concepts, etc. However, many systems approximate these semantic notions through the use of bigrams [1]–[3]. (ii) Given that the system can be applied to different summarization setups, we can apply it to summarization corpora of different genres and domains to explore the generalization power of different annotation types.

In order to evaluate the capability of the resulting system to estimate information importance, we propose two novel evaluation strategies, which aim at a direct evaluation of its information importance estimation capability instead of the final output of the summarization system. Thus, these measures allow us to focus on the key problem, and are independent from other components of the summarization system such as its ability to avoid redundancy or from variations in summary lengths.

II. Learning to estimate importance of text elements from pairwise preferences

Summarization systems can be grouped in two types of systems: abstractive and extractive systems. While abstractive models learn to write summaries from scratch, extractive models learn to extract sentences from the source documents and build summaries by concatenating the extracted sentences. We focus on extractive models in this paper since abstractive models usually do not use any linguistic annotations but produce summaries in an end-to-end manner. In particular, we employ a recently proposed summarization system called CPSum [1] as a basis for our experiments.

CPSum learns from pairwise preferences [4] to estimate the importance of text elements. A pairwise preference $o_i \succ o_j$ indicates that the summarization system should prefer to include

This work has been supported by the German Research Foundation (DFG) as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) under grant No. GRK 1994/1, and by a gift from NVIDIA Corporation.

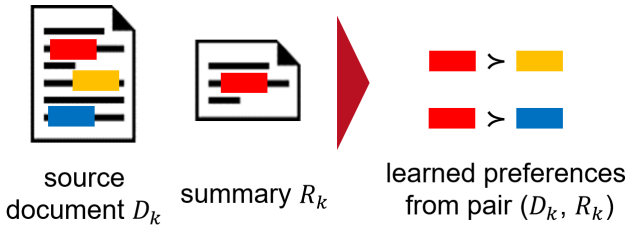

Fig. 1. Illustration of the learning process. Given many pairs of source documents and summaries (D_k, R_k) with annotated text elements, the system generates pairwise preferences. In this example, the system learns that the red element is preferred over the yellow and blue elements since the red element has been promoted to the summary whereas the yellow and the blue element do not appear in the summary.

text element o_i in the summary, rather than o_j. Ideally, the text elements represent semantic information nuggets, similar to summary content units(SCUs) in the Pyramid method [5]. Some information nuggets are more important than others. Consider, for example, the task of summarizing a transcript of a soccer match commentary. Information nuggets contained in the transcript are substitutions, fouls, passes, and goals. Even though fouls and passes occur much more frequently, the summary will primarily focus on the scored goals and the final result, since these are the most important information nuggets. This information can be learned from pairs of source documents (such as a transcripts) and sample summaries (D_k, R_k), where the system can generate pairwise preferences for any information item o_i that occurs in the summary over all other information items o_j in the document, i.e.,

$$o_i \succ o_j \Leftrightarrow o_i \in R_k, o_j \in D_k \setminus R_k. \tag{1}$$

In Figure 1 two preferences are created since the information sketched in red (mentioning e.g., a scored goal) occurs in the summary whereas the yellow and blue items (mentioning e.g., fouls and passes) do not occure in the summary.

CPSum learns from many such pairwise preferences. We apply the Bradley-Terry model [6] to learn the importance of individual text elements $o_i \in \mathcal{O}$ based on observed pairwise preferences. Given a set of pairwise preferences over elements o_i, the Bradley-Terry model estimates utility scores p_i such that

$$\Pr(o_i \succ o_j) = \frac{p_i}{p_i + p_j} \tag{2}$$

A unique maximum likelihood estimate for the utility scores can be found by the following algorithm [7]. First, we initialize all scores p_i with the uniform distribution $p_i \leftarrow 1/|\mathcal{O}|$, and then iteratively update the scores according to

$$p_i \leftarrow \alpha_i \cdot \frac{M_i}{\sum_{i \neq j} \frac{N_{ij}}{p_i + p_j}} \tag{3}$$

where M_i refers to the total number of preferences involving o_i as the preferred object, N_{ij} is the number of observations between elements o_i and o_j, and α_i is normalization factor so that $\sum_{j=1}^{|\mathcal{O}|} p_j = 1$. This update is repeated until the changes between two successive iterations become sufficiently small.

In [1], only bigrams have been considered as text elements o_i. In this work, we investigate the performance of a wide variety of linguistic annotation types with which we instantiate the set $\mathcal{O}$. We describe all annotation types used in Section IV.

III. Evaluating Importance Estimation

Computing similarities between system and reference summaries, as is typically done in summary evaluation with scores like ROUGE [8], has several disadvantages if we are mainly interested in evaluating the model's ability to estimate information importance: (i) ROUGE has to compute the semantic similarity between two texts, which is a very complex and yet unsolved problem. The predictions made by ROUGE are therefore noisy and not very accurate. (ii) ROUGE measures not only importance detection but also redundancy avoidance. Being able to evaluate both subtasks independently would provide us with better insights into the strengths and weaknesses of different approaches for the individual subtasks. (iii) Creating summaries is always performed with regard to a length restriction. The length parameter adds additional complexity to the evaluation and makes interpretation of results harder. We therefore propose two novel evaluation strategies which do not suffer from the three previously described limitations.

A. Evaluating via Ranking Prediction

Our first evaluation method evaluates if summarization systems are able to rank sentences according to their importance. We propose to use two rank correlation metrics to estimate the quality of the predicted ranking. The Kendall rank correlation coefficient (Kendall's tau) [9] computes the number of concordant pairs in the rankings. All disagreements are equally weighted, i.e., ranking mistakes in the bottom and in the top part of the ranking are equally penalized by Kendall's tau. Having a good agreement in the top part of the ranking is, however, more important than a good agreement in the bottom part of the ranking, since only highly ranked sentences will be included in a summary in practice. We therefore define a variant of the discounted cumulative gain (DCG) [10] which is frequently used in information retrieval. We define the *discounted cumulative ranking score* (DCRS) between two ranking functions π and $\hat{\pi}$ as

$$DCRS(\pi, \hat{\pi}) = \sum_{i=1}^{n} \frac{\frac{1}{\pi(o_i)}}{\ln(\hat{\pi}(o_i) + 1)}, \tag{4}$$

where $\pi(o_i)$ and $\hat{\pi}(o_i)$ indicate the rank of o_i according to π and $\hat{\pi}$, respectively. The difference to DCG is that the gain we use ($\frac{1}{\pi_f(o_i)}$) is only based on the rank and not on the utility of the elements. Following the most common variant of DCG, we also use a logarithmic discount factor [11]. In our experiments, we report results of the normalized DCRS (nDCRS) which maps all DCRS scores into $[0, 1]$ [12]. A random permutation of the list yields an nDCRS score of 0.5.

In addition to the rank correlation metrics, we also compute the precision@k score, which computes how many of the first k elements in the predicted ranking are contained in the first k elements of the target ranking.

B. Pairwise Preference Prediction

As a second evaluation method, we build two sets of sentences. Let set D contain all source sentences and let R contain all sentences from the reference summaries. We sample pairs of sentences d_i, r_i such that $d_i \in D$ and $r_i \in R$ and test the models' ability to distinguish good sentences stemming from a reference summary and bad sentences stemming from a source document (excluding sentences from the reference). The ordering of the sentences is randomized such that the models do not know which sentences have been drawn from D and R, respectively.

The models predict a preference label for each pair which indicates whether the first or the second sentence is better. Given n sampled pairs (d_i, r_i), we define the accuracy of a model's preference prediction ability as

$$acc = \frac{1}{n} \sum_{i=1}^{n} [[\hat{\pi}(r_i) > \hat{\pi}(d_i))]] \quad (5)$$

where $[[x]]$ denotes the indicator function mapping to 1 if x is true and 0 otherwise.

IV. Annotations under Investigation

In this section, we describe which types of annotations we investigate in this study, why we suspect that they could be helpful, and if/where downstream applicability has already been investigated. The annotation types are roughly organized according to increasing complexity.

Unigrams. The unigram annotation indicates if a given word type is present in the text. Uniframs were also used by [13] for importance estimation.

Bigrams. Bigrams capture consecutive bigrams appearing in the text and have been used before for estimating sentence importance in summarization systems [1]–[3], [14].

Trigrams. Trigrams are analogous to bigrams, but indicate the appearance of a consecutive sequence of three words. They are therefore able to capture longer phrases.

Verb Stems. For each verb in the text we use its lemma as a feature (e.g., *killing, killed* $\rightarrow$ *kill*). The intuition here is that particular verbs convey importance better than others. If a news article contains the information that someone has been killed, this information will most likely also be contained in the reference summary. On the other hand, it is often reported in news articles that person x said y (e.g. uttered an opinion) which might be a rather unimportant detail which is not contained in the summary.

Chunks. A recent study on interactive summarization [15] shows that chunks can also be used as an alternative to bigrams in a summarization system. Chunks are constituent parts of a sentence with a specific grammatical meaning (e.g. noun chunks, verb chunks). In this work we use the Treetagger chunker[1] and consider four chunk types, namely noun chunks (*NC*), verb chunks (*VC*), adverbial chunks (*ADVC*) and adjectival chunks (*ADJC*). As chunks capture grammatical meaning, we believe they are a viable replacement to bigrams and can capture richer importance features.

Named entities (NEs). This annotation type identifies mentions of entities and their semantic types, such as persons, locations, or organizations. We evaluate different variants of NE annotations. For example, the 21 entity types found by applying the CoreNLP named entity recognizer [16], e.g. *PERSON*, *CITY*, or *COUNTRY*; 91 fine-grained entity types from the FIGER type inventory [17], e.g. */person/politician* or */building/hotel*; and unique IDs for each entity. We obtain Freebase [18] entity IDs via an entity linking system [19] and then map these IDs to their FIGER type.

Frames. Following the theory of frame semantics [20], humans understand the meaning of words in terms of frames. FrameNet [21] provides an inventory of such frames which are used to provide a fine-grained interpretation of predicates in sentences by disambiguating the predicate's meaning with respect to frames. FrameNet annotations have been used for question answering and automatic text summarization [22], event detection [23], text understanding [24], and textual entailment [25]. To annotate all nouns and verbs of the texts with frames we use the neural network-based system of [26] which assigns a frame to a word based on the word itself and the surrounding context in the sentence.

Concepts. In order to identify concepts in text, we follow the work of [27] who primarily rely on open information extraction [28] to detect mentions of concepts and their relationships, and then use several measures of semantic similarity between them to cluster mentions of the same concept together. In this work, we use small concept clusters obtained by string matching different mentions (concepts string) and broader clusters based on semantic similarities (concepts sim). Compared to bigrams, which are directly defined on the lexical level, we expect the clustering to semantic groups to yield richer importance signals that generalize better.

Connotation frames (CFs). CFs are a new formalism for analyzing subjective roles and relationships implied by a given predicate [29]. For example, in *[Brazil]*$_{agent}$ *is suffering from [a failing economy]*$_{theme}$, the verb *suffering* indicates that the writer treats *Brazil* more sympathetically and the theme more as an "antagonist". *Brazil* most likely feels negatively towards the theme, it has been hurt, its "mental health" is distressed, but it is considered valuable. All these relationships are unified by a CF which contains labels for relationships inferable from the predicate *suffer*. Given that CFs capture implicit sentiment of the writer and sentiment between entities, we suspect that CFs can signal importance.

Discourse Relation (DR) Senses. DRs as annotated in PDTB [30] indicate specific thematic relations between clauses or sentences in discourse, such as *causation, contrast, or concession*. These relation *senses* can be explicitly marked (e.g., *but, whereas*) or are implicitly understood in unmarked juxtaposition of sentences. DRs capture a notion of *importance* at the level of text organization that is especially relevant for summarization. We use annotations from the output of the DR sense classification system of [31].

[1] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

TABLE I
RESULTS OF THE RANKING EXPERIMENTS

	Test on training data						Test on unseen data						Ablation experiment					
	Kendall's Tau		nDCRS		precision@k		Kendall's Tau		nDCRS		precision@k		Kendall's Tau		nDCRS		precision@k	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
bigram	**.504**	**.634**	**.536**	**.929**	**.536**	**.588**	**.306**	**.539**	**.253**	**.863**	**.253**	**.424**	**-.112**	**-.064**	**.124**	**.453**	**.124**	**.093**
cf-effect-object	-.049	.266	.085	.686	.085	.229	-.051	.269	.083	.687	.083	.230	-.092	-.016	.138	.487	.138	.115
cf-state-subject	-.044	.292	.088	.703	.088	.240	-.054	.284	.083	.697	.083	.234	-.092	-.016	**.137**	.487	**.137**	.116
chunk-concepts	**.343**	**.478**	**.363**	**.861**	**.363**	**.444**	**.175**	**.367**	**.206**	**.773**	**.206**	**.298**	**-.095**	**-.019**	.139	**.486**	.139	.115
concepts-string	**.181**	.276	**.259**	.699	**.259**	.263	.106	.193	.146	.639	.146	.179	-.087	-.012	.138	.490	.138	.115
concepts-sim	.130	.270	.168	.698	.168	.247	.093	.225	.135	.669	.135	.225	-.089	-.013	**.135**	.490	**.135**	**.114**
connotation-frames	-.006	.326	.091	.734	.091	.266	-.011	.335	.089	.739	.089	.267	-.092	-.017	.142	.487	.142	**.113**
entity-importance	-.065	-.027	.110	.532	.110	.194	-.076	-.060	.107	.510	.107	.193	-.089	-.007	.139	.495	.139	.119
entity-links	.187	.320	.232	.752	.232	.329	**.135**	.264	**.169**	.709	**.169**	.261	-.078	-.005	.142	.496	.142	.116
entity-type-coarse	.037	.087	.136	.577	.136	.157	.031	.100	.138	.582	.138	.155	-.087	-.004	.142	.497	.142	.117
entity-type-corenlp	.091	.352	.152	.758	.152	.315	.075	.358	.132	.766	.132	**.316**	**-.095**	**-.018**	**.136**	**.486**	**.136**	**.113**
entity-type-figer	.130	.275	.179	.714	.179	.248	.122	.272	.165	.709	.165	.243	-.083	-.007	.142	.494	.142	.115
entity-type-fine	.129	.274	.181	.713	.181	.252	.117	.269	.163	.708	.163	.236	-.083	-.007	.139	.494	.139	.115
FN-frames	.052	.399	.118	.785	.118	.317	.027	**.383**	.107	**.772**	.107	.297	**-.095**	**-.018**	.138	**.486**	.138	.115
FN-frames-nounsOnly	.154	**.487**	.168	**.848**	.168	**.402**	.116	**.474**	.133	**.836**	.133	**.364**	**-.095**	**-.018**	.139	**.485**	.139	.115
FN-frames-verbsOnly	.016	.216	.099	.646	.099	.186	.010	.209	.096	.639	.096	.186	-.093	-.017	.138	**.486**	.138	**.114**
sentiment-annos	.097	.230	.209	.693	.209	.263	.068	.215	.148	.673	.148	.222	-.084	-.010	.144	.492	.144	.117
discours-rel	.009	.232	.134	.644	.134	.173	.011	.234	.133	.646	.133	.174	-.092	-.016	.139	.488	.139	.117
trigram	**.314**	**.535**	**.443**	**.866**	**.443**	**.426**	**.172**	.366	**.186**	.760	**.186**	.241	**-.095**	-.017	**.137**	**.486**	**.137**	**.114**
unigram	**.401**	**.693**	**.350**	**.931**	**.350**	**.568**	**.300**	**.654**	**.260**	**.913**	**.260**	**.515**	-.093	-.017	.141	.487	.141	.116
verb-stem	.088	.275	.147	.702	.147	.251	.042	.250	.114	.671	.114	.215	-.093	-.017	.138	**.486**	.138	.115

V. EXPERIMENTS

We performed a wide range of experiments to evaluate which linguistic annotations are most helpful for the used model. We first describe the used data in Section V-A before we describe ranking experiments in SectionV-B and preference label prediction experiments in Section V-C.

A. Data

For the experiment, we use three well-known multi-document summarization datasets, namely the DUC 2004 (DUC2004), TAC 2008 (TAC2008), and TAC 2009 (TAC2009) corpora. All corpora are freely available upon request[2] which allows a better reproducibility of our experiments. Each corpus contains about 50 summarization topics which is a collection of 10 related source documents. For each document topic, human annotators created 4 reference summaries.

We estimate sentence scores based on the text element utilities defined in Section II depending on whether ROUGE recall or ROUGE precision has to be predicted. For ROUGE recall, we simply add the utility scores of all text elements that appear in a sentence to compute the sentence utility score. If I is the list of all element indices that appear in a sentence s, we define the utility score of s to be $\sum_{i \in I} p_i$. For ROUGE precision, we divide the score for ROUGE recall by the length of the sentence: $\frac{1}{|s|} \sum_{i \in I} p_i$.

B. Ranking Experiments

We provide the results of three different ranking experiments according to Section III-A in Table I. To generate target rankings, we extract all sentences from the source documents and rank the sentences according to their ROUGE recall and ROUGE precision. [32] found that summarization systems with a greedy sentence selection strategy perform well if the systems are able to rank sentences according to their ROUGE precision scores. ILP-based systems perform well if sentences can be ranked according to ROUGE recall scores. We therefore use both scores to generate target rankings. Results for ROUGE precision and ROUGE recall are in columns P and R, respectively. We report Kendall's Tau, nDCRS, and precision@k (with $k = 20$ in all experiments) scores for many different annotation types including bigrams which were used in the original study [1]. For better visualization, we highlight the best 5 results in every column. For the first two experiments (left and middle), higher scores are better. For the ablation experiments (right), lower scores are better. Details and analysis of the experiments are provided in the next three subsections.

1) Which Annotation Types Can Potentially Convey Importance?: In the first experiment, we investigate whether the annotations types can potentially be used by the model to learn about the importance of information. To this end, we use in this experiment the same data for training and testing. If the model is able to learn based on the annotations, it should be able to achieve a reasonable performance in this setting. We provide the results of this experiment in the left part "Test on training data" in Table I.

The best performing annotation types are unigrams, bigrams, trigrams, and chunk-concepts which is not surprising in this experimental setup since we test the performance on the training data. This means that annotations which are close to the text are able to adapt well to the data. Similarly, chunk-concepts and concepts-string perform reasonably well. Very surprising is the large performance difference of FN-frames-nounsOnly between ROUGE precision and ROUGE recall ranking prediction (columns P and R). FN-frames-nounsOnly annotations do not perform well for ROUGE precision, but are among the best annotations for ROUGE recall. We observe that cf-effect-object and cf-state-subject perform worst. This can be explained by the fact that these annotations occur

[2]http://duc.nist.gov and https://tac.nist.gov

very frequently in both source documents and summaries. The model is therefore not able to use these annotations as importance indicators. We annotated all text elements automatically, meaning that higher-level annotations might be inaccurate to some extent. This might be another explanation for the superior performance of low-level annotations.

2) Which Annotation Types Convey Importance Across Topics?: In this experiment we analyze how well the model is able to transfer learned knowledge to other, unseen topics. The results are presented in the middle "Test on unseen data" of Table I. From the four datasets used in this paper, we perform four experiments in which we select one dataset as test set and use the remaining three datasets for training. We report the average performance of the four experiments.

The best three performing annotations are unigrams, bigrams, and chunk-concepts. Compared to the first experiments, we observe that trigrams lost performance. The reason is that it is likely that there are more, unseen trigrams in the test data for which the model was not able to learn importance scores. The ability to generalize of trigrams is limited. Somewhat surprising is that chunk-concepts, which are also very close to the surface text, perform still quite well. chunk-concepts seem to generalize better than trigrams. FN-frames-nounsOnly did not loose much performance compared to the first experiment. This indicates that FN-frames-nounsOnly does not overfit to the training data and generalizes well. Entity-links, entity-type-corenlp, entity-type-fine and FN-frames are also among the top 5 in some columns. We also see a rather large relative performance drop of bigrams compared to, for example, unigrams.

3) Ablation Experiments: In the last ranking experiment, we perform an ablation study within experimental setting 2. We aggregate the rankings of all annotation types except of one. The first line on the right part in Table I, for example, contains the aggregated ranking of all annotation elements except of bigrams. The scores indicate how much performance is lost if a particular annotation element is removed from the ensemble. Lower scores are therefore better. As aggregation function, we simply compute the average rank for each sentence and rank the sentence according to the averaged ranks.

The biggest drop in performance is observed when bigrams or entity-type-corenlp are removed from the ensemble. entity-type-corenlp annotation seems to contribute to the ensemble even though they did not perform well in the first two experiments. We observe that unigrams do not contribute to the ensemble even though they showed a very good performance in the first two experiments. Similarly to the second experiment, frame-based annotations show reasonably good performance. Annotations based on connotation frames also rank among the top 5.

C. Predicting Preference Labels for Source And Reference Sentences

In the next experiment, we show the evaluation of the different annotation types based on our second novel evaluation method described in Section III-B. With this evaluation method, we test how well a model can distinguish between sentences sampled from source documents and summaries. The results of the experiment are displayed in Table II. We use 3 of 4 datasets for training and test on the remaining dataset.

TABLE II
RESULTS OF THE PAIRWISE PREFERENCE EXPERIMENTS.

	DUC 2003	DUC 2004	TAC 2008	TAC 2009	average
bigram	**0.573**	0.538	0.415	0.445	0.493
cf-effect-object	0.538	0.520	**0.663**	**0.743**	**0.616**
cf-state-subject	0.548	0.439	0.420	0.512	0.480
chunk-concepts	**0.641**	**0.613**	**0.556**	0.602	**0.603**
concepts-string	0.513	0.429	0.371	0.382	0.424
concepts-sim	0.520	0.468	0.438	0.473	0.475
connotation-frames	0.551	0.556	0.546	0.592	0.561
entity-importance	**0.597**	**0.634**	**0.655**	**0.658**	**0.636**
entity-links	0.510	0.450	0.370	0.364	0.424
entity-type-coarse	0.512	0.487	**0.664**	**0.695**	**0.590**
entity-type-corenlp	**0.582**	**0.608**	0.551	**0.616**	0.589
entity-type-figer	0.495	0.487	0.453	0.408	0.461
entity-type-fine	0.497	0.490	0.456	0.405	0.462
FN-frames	0.474	0.497	0.515	0.496	0.496
FN-frames-nounsOnly	0.521	0.537	0.531	0.539	0.532
FN-frames-verbsOnly	0.490	0.487	0.468	0.507	0.488
sentiment-annos	0.430	0.402	0.353	0.356	0.385
discours-rel	0.550	**0.608**	**0.628**	**0.604**	**0.598**
trigram	0.373	0.285	0.210	0.254	0.281
unigram	**0.617**	**0.601**	0.530	0.553	0.575
verb-stem	0.497	0.517	0.515	0.500	0.507

The results are very different compared to the ranking results in Table I. The model is best able to use entity-importance to distinguish between source and summary sentences, followed by cf-effect-object, chunk-concepts, and now also discourse-rel performs consistently well. Bigrams, which performed very well in the ranking experiments, performs poorly in this experiment. They show a bad performance in particular in the TAC 2008 and TAC 2009 corpora. cf-effect-object and entity-type-coarse perform well in the TAC datasets. entity-importance does not have the highest performance in the TAC datasets but also works well in the DUC datasets which leads in the end to the overall best performance.

VI. CONCLUSIONS

We studied if and how well a wide range of linguistic annotations can improve the performance of a complex high-level task, namely the detection of important information.

Our ranking experiments show that annotations that are close to the surface text such as n-grams and chunks perform best. Researchers should therefore consider these annotations while developing summarization systems. They can also serve as simple annotations to build strong baselines. However, other annotations also showed potential in specific situations. In particular entity and frame annotations are able to improve the performance in some cases. In our pairwise preference prediction experiments, we observed a different behavior. Bigrams, which performed well for ranking, did not perform well in this experiment. Instead, entity-based annotations, connotation information and discourse relations perform well in distinguishing source from reference sentences.

In conclusion, this study provided us with insights about a wide range of possible annotation types. We showed that simple bigrams perform well in many experiments. However, more complex annotations also showed good performance, in particular in the pairwise preference prediction experiments.

VII. Future Work

For future work, we would be curious if and how much similar annotations can improve the performance of other summarization systems. The model we used is based on symbolic representations and is not designed to generalize across different annotation elements based on a feature representation of the annotations. Implementing a soft-matching for the preferences could be very promising to generalize from observed to unobserved annotation elements if features such as word embeddings could be used in the summarization system. Furthermore, we would be interested if similar observations can be made in non-newswire [33], [34] and in non-English corpora.

References

[1] M. Zopf, E. Loza Mencía, and J. Fürnkranz, "Beyond Centrality and Structural Features: Learning Information Importance for Text Summarization," in *Proceedings of the 20th Conference on Computational Natural Language Learning*, pp. 84–94, 2016.

[2] D. Gillick, B. Favre, D. Hakkani-Tür, B. Bohnet, Y. Liu, and S. Xie, "The ICSI/UTD Summarization System at TAC 2009," in *Proceedings of the Second Text Analysis Conference*, 2009.

[3] H. Lin and J. Bilmes, "A Class of Submodular Functions for Document Summarization," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 510–520, 2011.

[4] J. Fürnkranz and E. Hüllermeier, eds., *Preference Learning*. Springer-Verlag, 2010.

[5] A. Nenkova, R. J. Passonneau, and K. R. McKeown, "The pyramid method: Incorporating human content selection variation in summarization evaluation," *ACM Transactions on Speech and Language Processing*, vol. 4, no. 2, p. 4, 2007.

[6] R. A. Bradley and M. E. Terry, "Rank Analysis of Incomplete Block Designs," *Biometrika*, vol. 39, no. 3, pp. 324–345, 1952.

[7] E. Zermelo, "Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, vol. 29, pp. 436–460, 1929.

[8] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, pp. 25–26, 2004.

[9] M. G. Kendall, "A New Measure of Rank Correlation," *Biometrika*, vol. 30, no. 1/2, p. 81, 1938.

[10] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[11] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, and T.-Y. Liu, "A Theoretical Analysis of NDCG Ranking Measures," *Proceedings of the 26th Annual Conference on Learning Theory*, pp. 1–30, 2013.

[12] W. Chen, T.-y. Liu, Y. Lan, Z.-m. Ma, and H. Li, "Ranking Measures and Loss Functions in Learning to Rank.," *Advances in Neural Information Processing Systems 22*, pp. 315–232, 2009.

[13] K. Hong and A. Nenkova, "Improving the Estimation of Word Importance for News Multi-Document Summarization.," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 712–721, 2014.

[14] J. Carbonell and J. Goldstein, "The use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," in *Proceedings of the 21st Annual ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR 1998)*, pp. 335–336, ACM, 1998.

[15] A. P.V.S. and C. M. Meyer, "Joint optimization of user-desired content in multi-document summaries by learning from user feedback," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. Volume 1: Long Paper, pp. 1353–1363, 2017.

[16] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics System Demonstrations*, pp. 55–60, 2014.

[17] X. Ling and D. S. Weld, "Fine-grained entity recognition.," in *AAAI*, 2012.

[18] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data,* Vancouver, B.C., Canada, 10–12 June 2008, pp. 1247–1250, 2008.

[19] B. Heinzerling, A. Judea, and M. Strube, "HITS at TAC KBP 2015: Entity discovery and linking, and event nugget detection," in *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, 2015.

[20] C. J. Fillmore, "Frame Semantics and the Nature of Language," *Annals of the New York Academy of Sciences*, vol. 280, no. 1, pp. 20–32, 1976.

[21] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet project," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, (Stroudsburg, PA, USA), pp. 86–90, 1998.

[22] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Comput. Linguist.*, vol. 28, no. 3, pp. 245–288, 2002.

[23] E. Spiliopoulou, E. Hovy, and T. Mitamura, "Event detection using frame-semantic parser," in *Proceedings of the Events and Stories in the News Workshop*, pp. 15–20, 2017.

[24] C. J. Fillmore and C. F. Baker, "Frame Semantics for Text Understanding," in *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, 2001.

[25] R. B. Aharon, I. Szpektor, and I. Dagan, "Generating Entailment Rules from FrameNet," in *Proceedings of the ACL 2010 Conference Short Papers*, pp. 241–246, Association for Computational Linguistics, 2010.

[26] S. Hartmann, I. Kuznetsov, T. Martin, and I. Gurevych, "Out-of-domain FrameNet semantic role labeling," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 471–482, Apr. 2017.

[27] T. Falke, C. M. Meyer, and I. Gurevych, "Concept-Map-Based Multi-Document Summarization using Concept Coreference Resolution and Global Importance Optimization," in *Proceedings of the 8th International Joint Conference on Natural Language Processing*, (Taipei, Taiwan), pp. 801–811, 2017.

[28] Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni, "Open Language Learning for Information Extraction," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (Jeju Island, Korea), pp. 523–534, 2012.

[29] H. Rashkin, S. Singh, and Y. Choi, "Connotation frames: A data-driven investigation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, (Berlin, Germany), pp. 311–321, Association for Computational Linguistics, August 2016.

[30] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The Penn Discourse treebank 2.0.," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, (Marrakech, Morocco), 2008.

[31] T. Mihaylov and A. Frank, "Discourse relation sense classification using cross-argument semantic similarity based on word embeddings," in *Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task*, Aug. 2016.

[32] M. Zopf, E. Loza Mencía, and J. Fürnkranz, "Which Scores to Predict in Sentence Regression for Text Summarization?," in *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1782–1791, 2018.

[33] M. Zopf, M. Peyrard, and J. Eckle-Kohler, "The Next Step for Multi-Document Summarization : A Heterogeneous Multi-Genre Corpus Built with a Novel Construction Approach," in *Proceedings of the 26th International Conference on Computational Linguistics*, pp. 1535–1545, 2016.

[34] M. Zopf, "auto-hMDS: Automatic Construction of a Large Heterogeneous Multi-Document Summarization Corpus," in *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pp. 3228–3233, 2018.