Overview of the Author Identification Task at PAN 2013

Patrick Juola & Efstathios Stamatatos

Duquesne University University of the Aegean

Outline

- Task definition
- Evaluation setup
- Evaluation corpus
- Performance measures
- Results
- Survey of approaches
- Conclusions

Author Identification Tasks

- Closed-set: there are several candidate authors, each represented by a set of training data, and one of these candidate authors is assumed to be the author of unknown document(s)
- Open-set: the set of potential authors is an open class, and "none of the above" is a potential answer
- Authorship verification: the set of candidate authors is a singleton and either he wrote the unknown document(s) or "someone else" did

Fundamental Problems

- *Given two documents, are they by the same author?* [Koppel *et al.*, 2012]
- *Given a set of documents (no more than 10, possibly only one) by the same author, is an additional (out-of-set) document also by that author?*

- Every authorship attribution case can be broken down into a set of such problems

Evaluation Setup

- One problem comprises a set of documents of known authorship by the same author and exactly one document of questioned authorship
- All the documents within a problem are matched in language, genre, theme, and date of writing
- Participants were asked to produce a binary yes/no answer and, optionally, a confidence score:
  - a real number in the set [0,1] inclusive, where 1.0 corresponds to “yes” and 0.0 corresponds to “no”
- Any problem could be left unanswered
- Software submissions were required
- Early-bird evaluation was supported

Evaluation Corpus

- English, Greek, and Spanish are covered
- Language information is encoded in the problem labels
- The distribution of positive and negative problems (in every language-specific sub-corpus) was balanced
- Problems per corpus/language:

Corpus	English	Greek	Spanish
Training	10	20	5
(Early-bird evaluation)	(20)	(20)	(15)
Final evaluation	30	30	25
Total	40	50	30

English Part of the Corpus

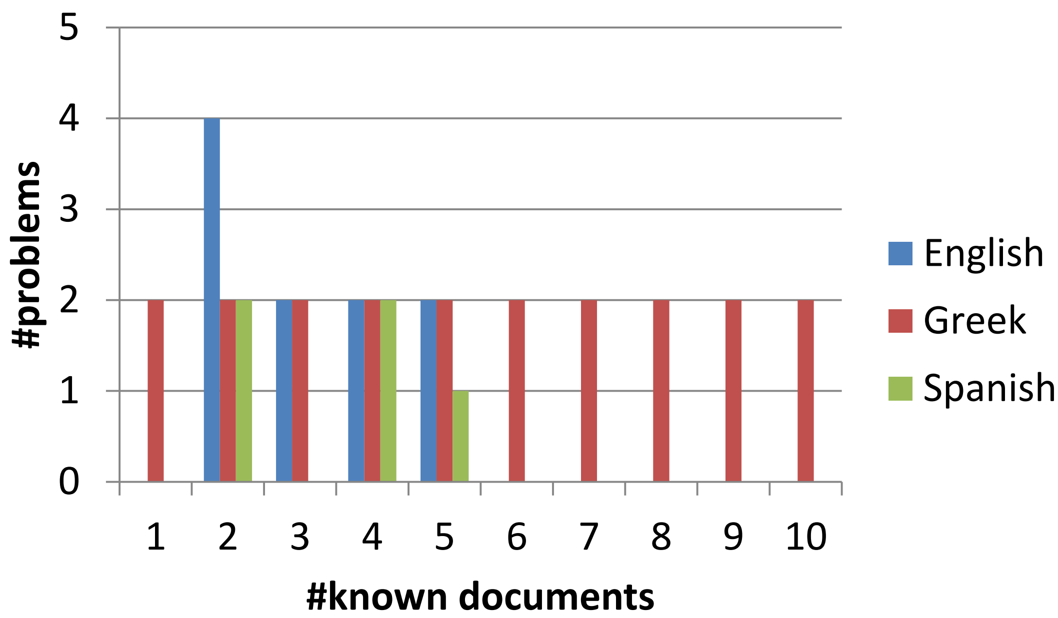
- Collected by Patrick Brennan of *Juola & Associates*
- Consists of extracts from published textbooks on computer science and related disciplines, culled from an on-line repository
  - A relatively controlled universe of discourse
  - A relatively unstudied genre
- A pool of 16 authors was selected and their works were collected
- Each document was around 1,000 words each and collected by hand from the larger works
- Formulas and computer code was removed
- Some of the paired documents are members of a very narrow genre
  - e.g. textbooks regarding Java programming
- Others are more divergent
  - e.g. Cyber Crime vs. Digital Systems Design)

Greek Part of the Corpus

- Comprises newspaper articles published in the Greek weekly newspaper TO BHMA from 1996 to 2012
- A pool of more than 800 opinion articles by about 100 authors was downloaded
- The length of each article is at least 1,000 words
- All HTML tags, scripts, title/subtitles of the article and author names were removed semi-automatically
- In each verification problem, texts with strong thematic similarities indicated by the occurrence of certain keywords
- To make the task more challenging, a stylometric analysis [Stamatatos, 2007] was used to detect stylistically similar or dissimilar documents
  - In problems where the true answer is positive the unknown document was selected to have relatively low similarity from the other known documents
  - When the true answer is negative, the unknown document (by a certain author) was selected to have relatively low dissimilarity from the known documents (by another author)

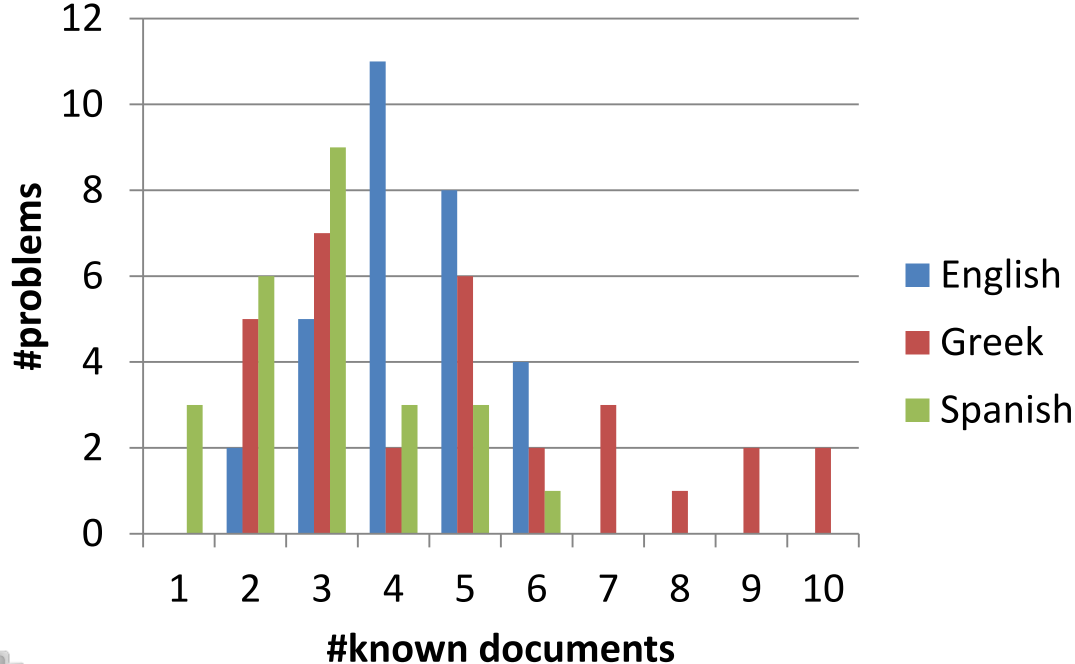
Spanish Part of the Corpus

- Collected in part by Sheila Queralt of Universitat Pompeu Fabra and by Angela Melendez of Duquesne University
- Consisted of excerpts from newspaper editorials and short fiction



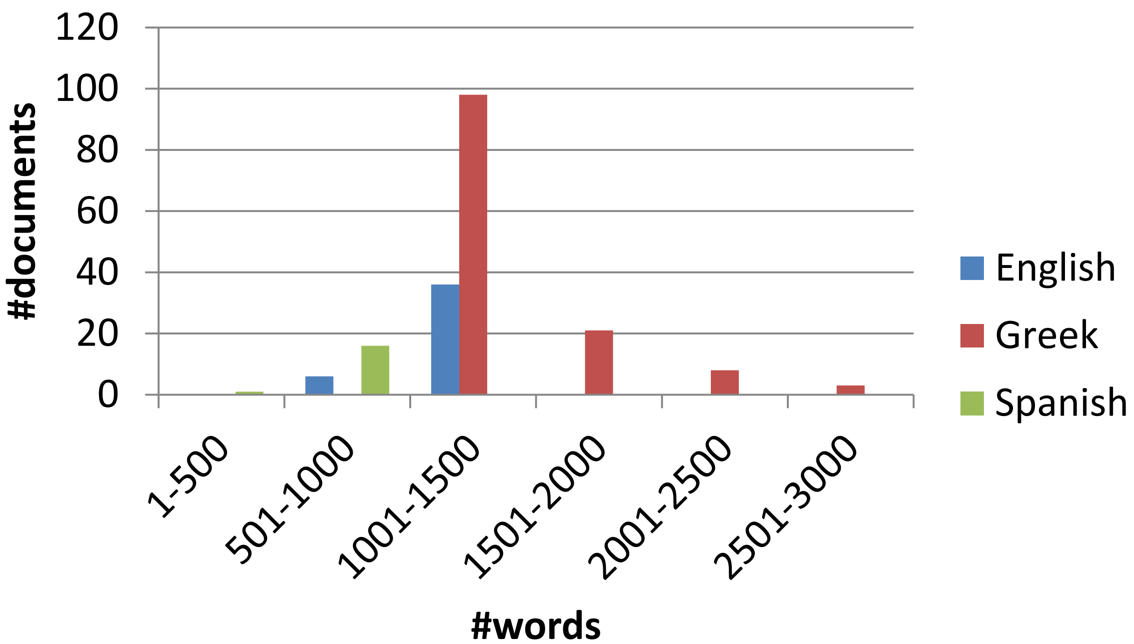
Training corpus

Distribution of known documents over the problems

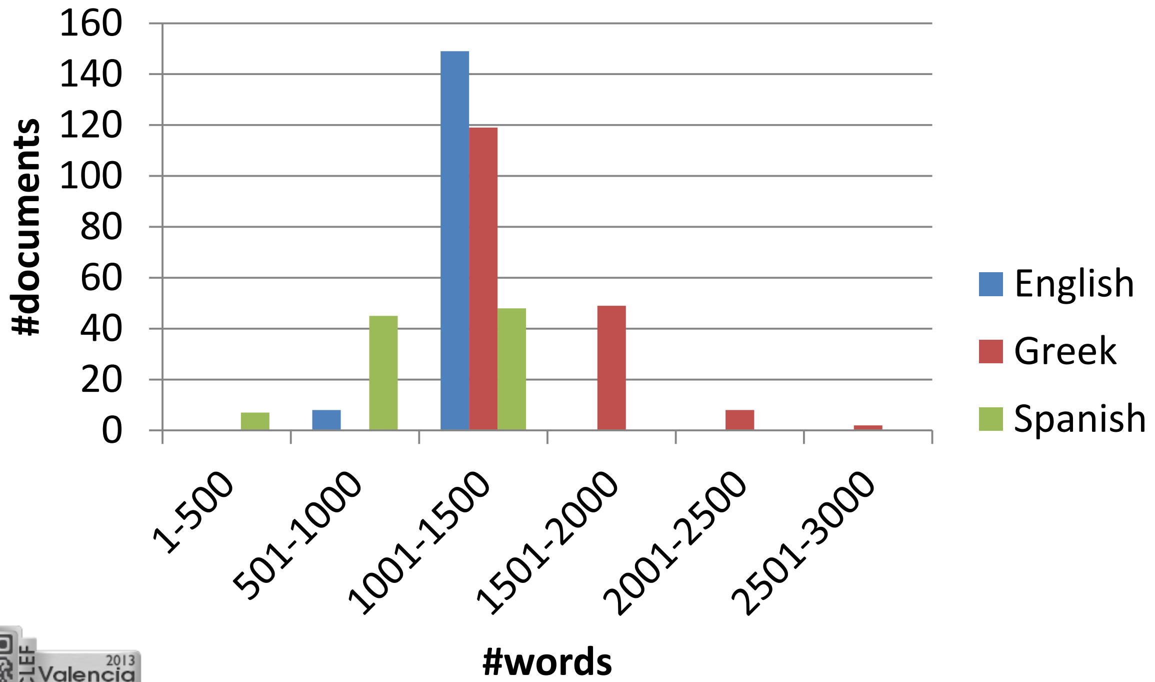


Evaluation corpus

Text-length distribution



Training corpus



Evaluation corpus

Performance Measures

- Overall results and results per language
- Binary yes/no answers:
  - *Recall = #correct_answers / #problems*
  - *Precision = #correct_answers / #answers*
  - F_1 (used for final ranking)
- Real scores:
  - ROC-AUC
- Runtime

Submissions

- 18 software submissions
  - From Australia, Austria, Canada (2), Estonia, Germany (2), India, Iran, Ireland, Israel, Mexico (2), Moldova, Netherlands (2), Romania, UK
- 16 notebook submissions
- 8 teams used the early-bird evaluation phase
- 9 teams produced both binary answers and real scores

Overall Results

Rank	Submission	F_1	Precision	Recall	Runtime
1	Seidman	**0.753**	0.753	**0.753**	65476823
2	Halvani et al.	0.718	0.718	0.718	8362
3	Layton et al.	0.671	0.671	0.671	9483
3	Petmanson	0.671	0.671	0.671	36214445
5	Jankowska et al.	0.659	0.659	0.659	240335
5	Vilariño et al.	0.659	0.659	0.659	5577420
7	Bobicev	0.655	0.663	0.647	1713966
8	Feng&Hirst	0.647	0.647	0.647	84413233
9	Ledesma et al.	0.612	0.612	0.612	32608
10	Ghaeini	0.606	0.671	0.553	125655
11	van Dam	0.600	0.600	0.600	9461
11	Moreau&Vogel	0.600	0.600	0.600	7798010
13	Jayapal&Goswami	0.576	0.576	0.576	7008
14	Grozea	0.553	0.553	0.553	406755
15	Vartapetiance&Gillam	0.541	0.541	0.541	419495
16	Kern	0.529	0.529	0.529	624366
	BASELINE	0.500	0.500	0.500	
17	Veenman&Li	0.417	**0.800**	0.282	962598
18	Sorin	0.331	0.633	0.224	3643942

Results for English

Submission	F_1	Precision	Recall
Seidman	**0.800**	**0.800**	**0.800**
Veenman&Li	**0.800**	**0.800**	**0.800**
Layton et al.	0.767	0.767	0.767
Moreau&Vogel	0.767	0.767	0.767
Jankowska et al.	0.733	0.733	0.733
Vilariño et al.	0.733	0.733	0.733
Halvani et al.	0.700	0.700	0.700
Feng&Hirst	0.700	0.700	0.700
Ghaeini	0.691	0.760	0.633
Petmanson	0.667	0.667	0.667
Bobicev	0.644	0.655	0.633
Sorin	0.633	0.633	0.633
van Dam	0.600	0.600	0.600
Jayapal&Goswami	0.600	0.600	0.600
Kern	0.533	0.533	0.533
BASELINE	0.500	0.500	0.500
Vartapetiance&Gillam	0.500	0.500	0.500
Ledesma et al.	0.467	0.467	0.467
Grozea	0.400	0.400	0.400

Results for Greek

Submission	F_1	Precision	Recall
Seidman	**0.833**	**0.833**	**0.833**
Bobicev	0.712	0.724	0.700
Vilariño et al.	0.667	0.667	0.667
Ledesma et al.	0.667	0.667	0.667
Halvani et al.	0.633	0.633	0.633
Jayapal&Goswami	0.633	0.633	0.633
Grozea	0.600	0.600	0.600
Jankowska et al.	0.600	0.600	0.600
Feng&Hirst	0.567	0.567	0.567
Petmanson	0.567	0.567	0.567
Vartapetiance&Gillam	0.533	0.533	0.533
BASELINE	0.500	0.500	0.500
Kern	0.500	0.500	0.500
Layton et al.	0.500	0.500	0.500
van Dam	0.467	0.467	0.467
Ghaeini	0.461	0.545	0.400
Moreau&Vogel	0.433	0.433	0.433
Sorin	-	-	-
Veenman&Li	-	-	-

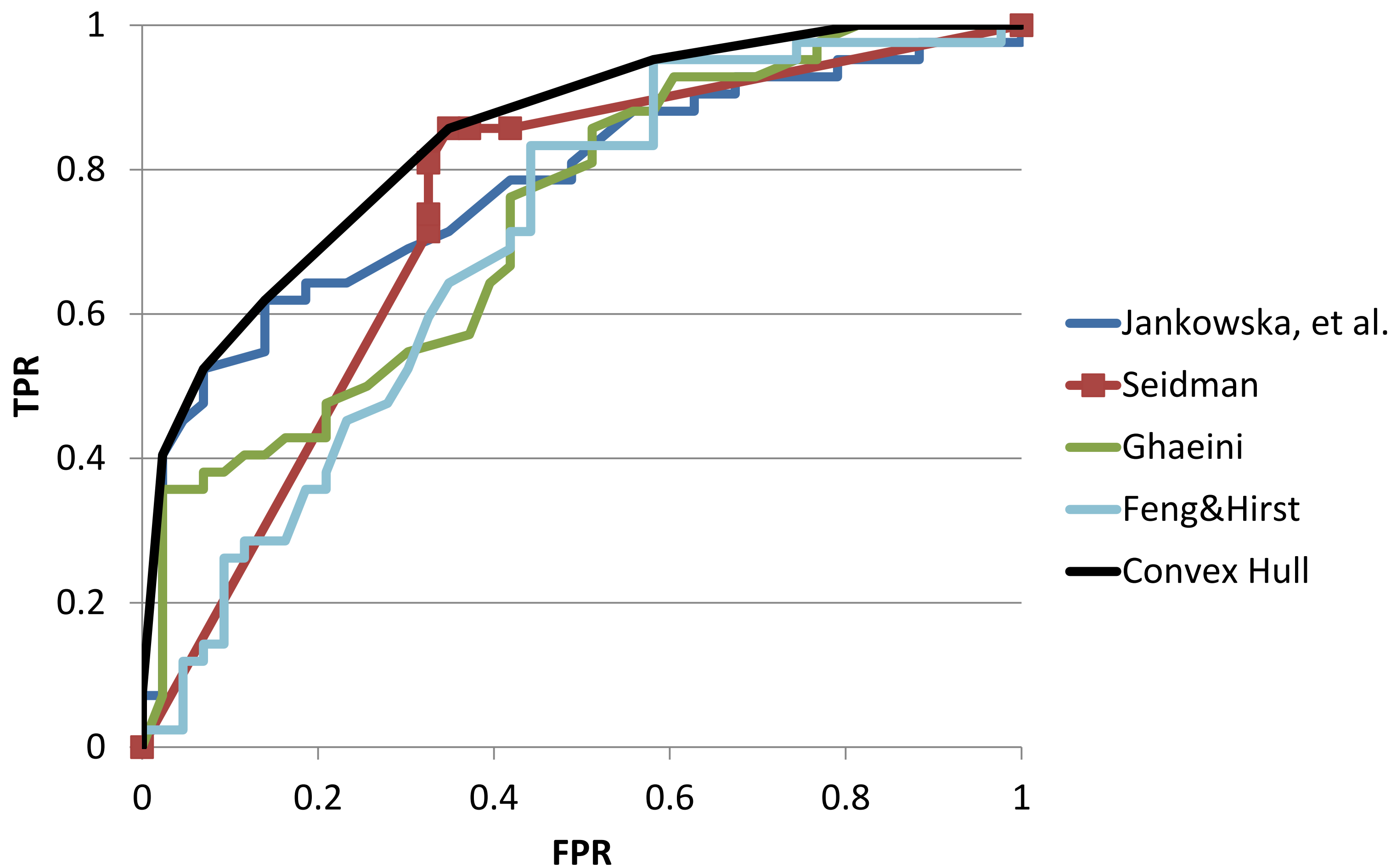
Results for Spanish

Submission	F_1	Precision	Recall
Halvani et al.	**0.840**	**0.840**	**0.840**
Petmanson	0.800	0.800	0.800
Layton et al.	0.760	0.760	0.760
van Dam	0.760	0.760	0.760
Ledesma et al.	0.720	0.720	0.720
Grozea	0.680	0.680	0.680
Feng&Hirst	0.680	0.680	0.680
Ghaeini	0.667	0.696	0.640
Jankowska et al.	0.640	0.640	0.640
Bobicev	0.600	0.600	0.600
Moreau&Vogel	0.600	0.600	0.600
Seidman	0.600	0.600	0.600
Vartapetiance&Gillam	0.600	0.600	0.600
Kern	0.560	0.560	0.560
Vilariño et al.	0.560	0.560	0.560
BASELINE	0.500	0.500	0.500
Jayapal&Goswami	0.480	0.480	0.480
Sorin	-	-	-
Veenman&Li	-	-	-

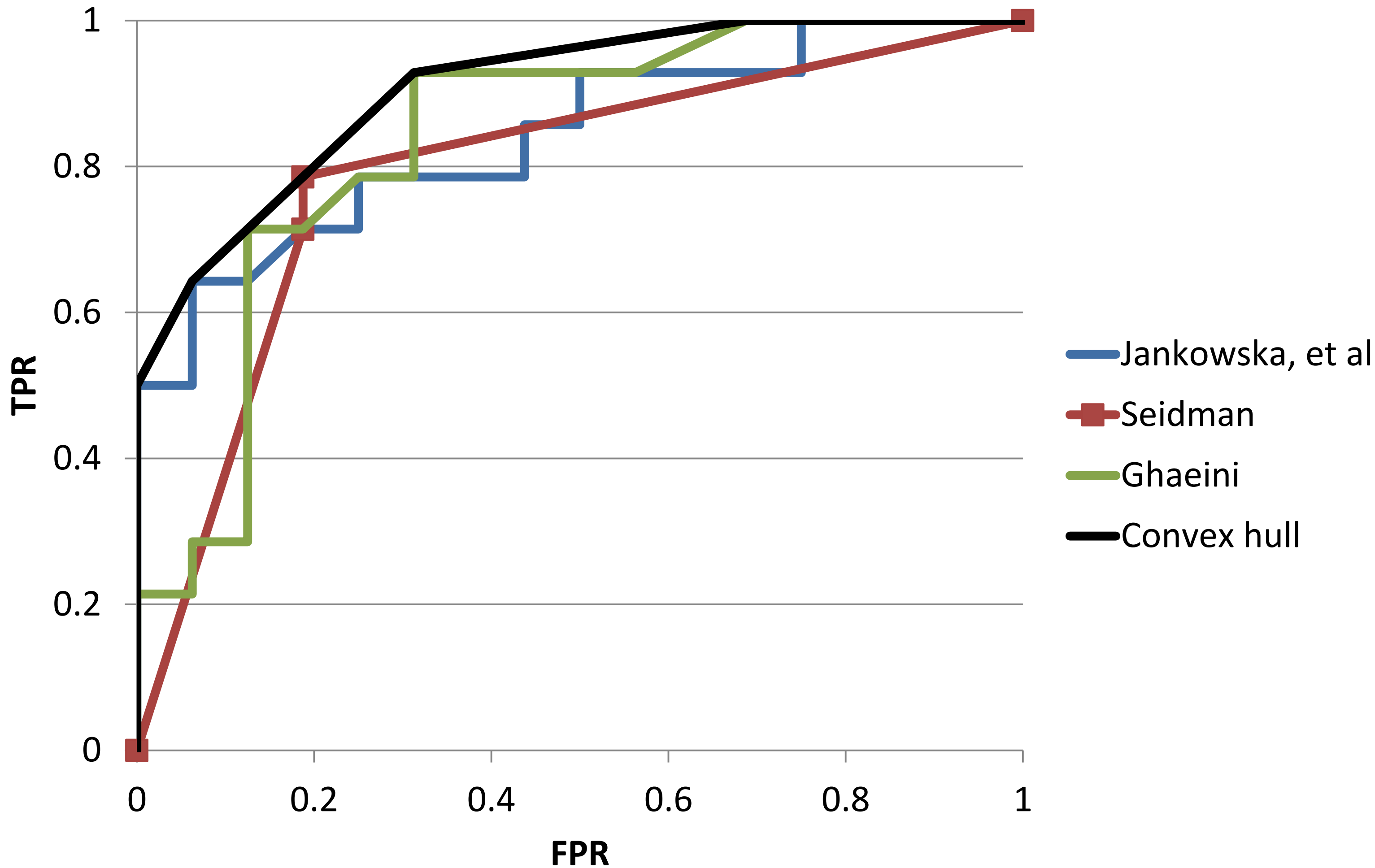
Overall Results (ROC-AUC)

Rank	Submission	Overall	English	Greek	Spanish
1	Jankowska, et al.	**0.777**	**0.842**	0.711	0.804
2	Seidman	0.735	0.792	**0.824**	0.583
3	Ghaeini	0.729	0.837	0.527	**0.926**
4	Feng&Hirst	0.697	0.750	0.580	0.772
5	Petmanson	0.651	0.672	0.513	0.788
6	Bobicev	0.642	0.585	0.667	0.654
7	Grozea	0.552	0.342	0.642	0.689
	BASELINE	0.500	0.500	0.500	0.500
8	Kern	0.426	0.384	0.502	0.372
9	Layton et al.	0.388	0.277	0.456	0.429

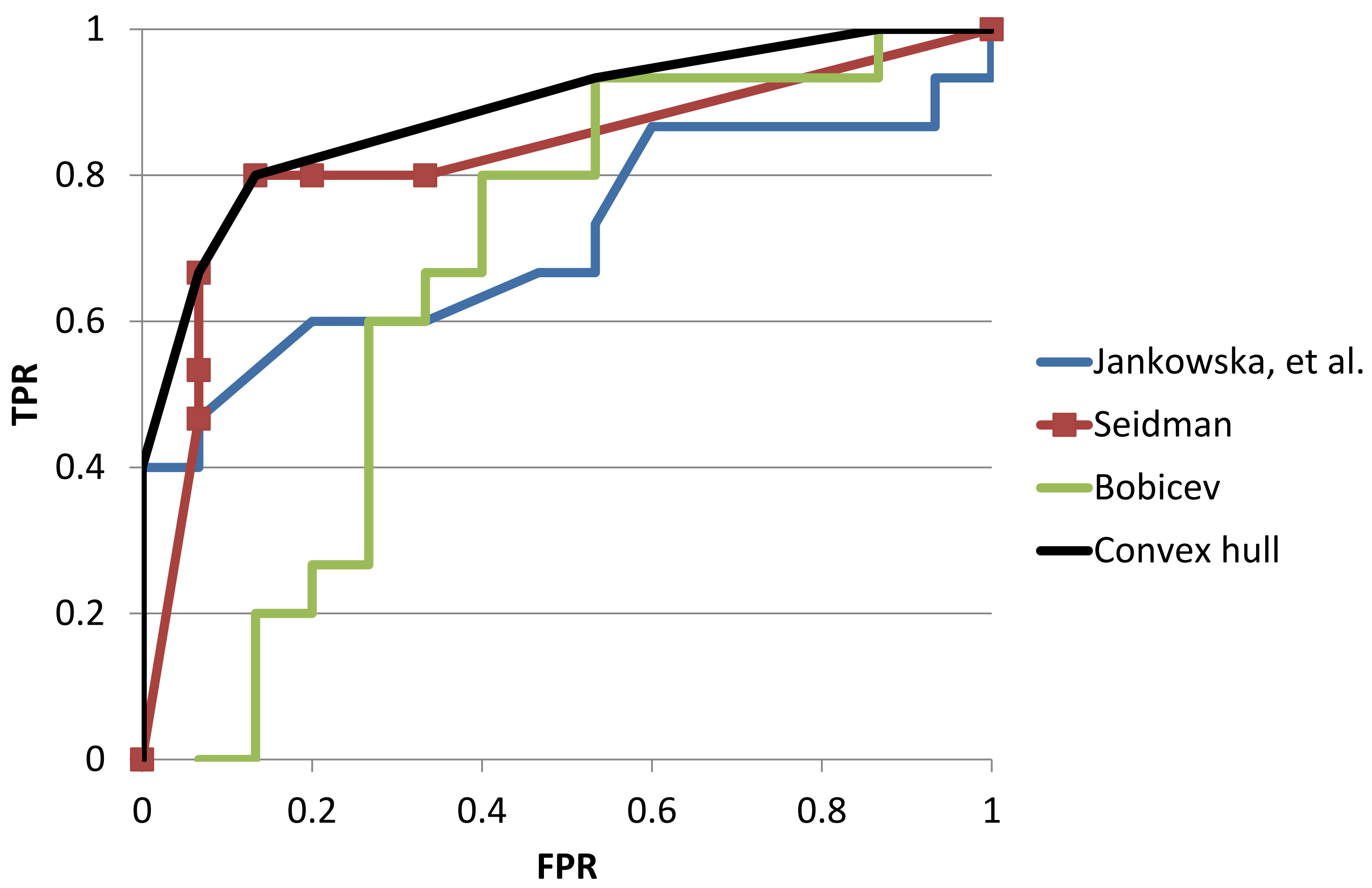
Overall Results (ROC)



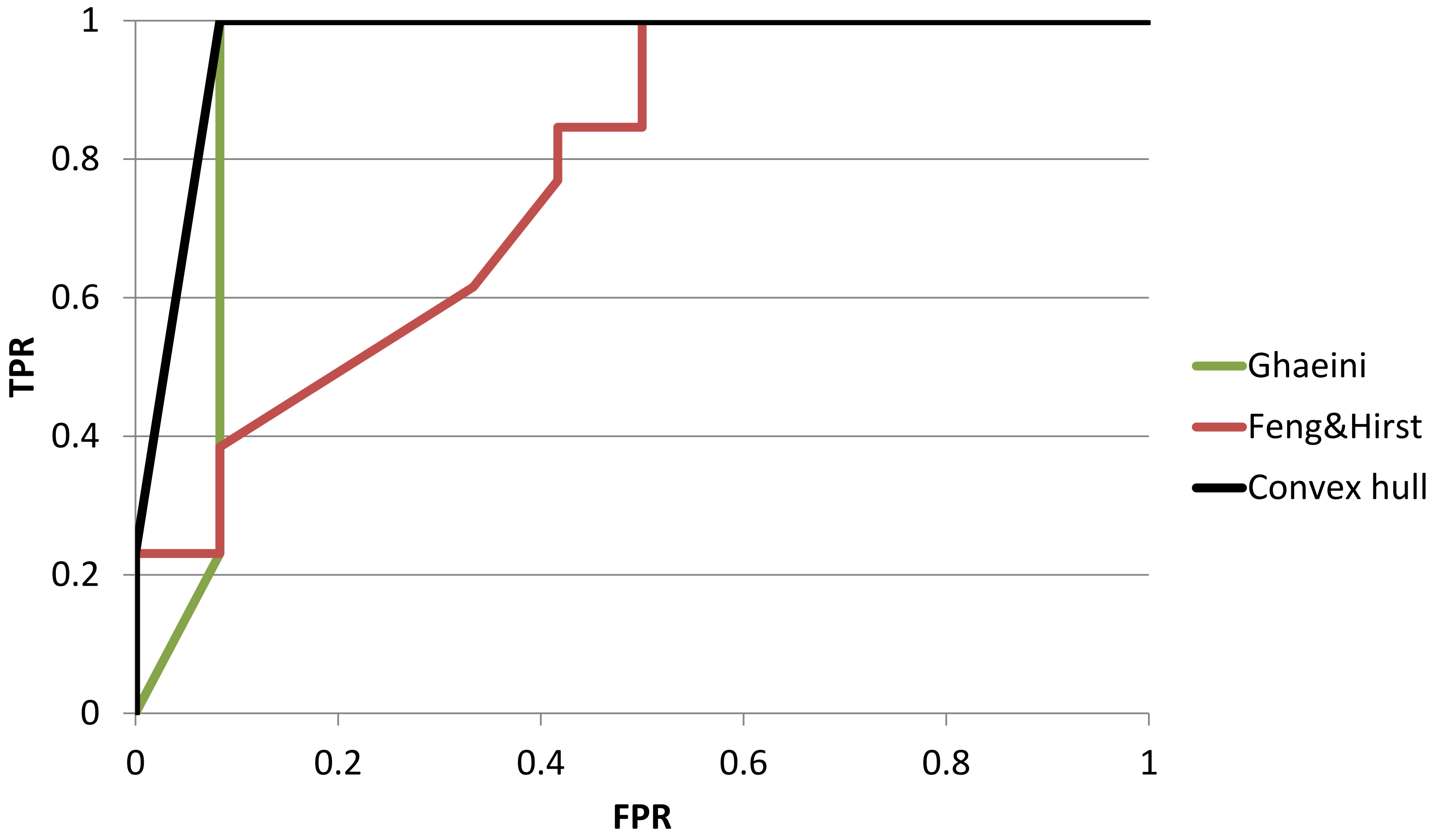
Results for English (ROC)



Results for Greek (ROC)



Results for Spanish (ROC)



Early-bird Evaluation

- To help participants build their approaches in time
  - Early detection and fix of bugs
- To provide an idea of the effectiveness on a part of the evaluation corpus
- In total, 8 teams used this option

Early-bird vs. Final Evaluation

Submission	Early-bird	Final	Difference
Jankowska, et al.	0.720	0.659	**-0.061**
Layton, et al.	0.680	0.671	**-0.009**
Halvani, et al.	0.660	0.718	**0.058**
Ledesma, et al.	0.620	0.612	**-0.008**
Jayapal&Goswami	0.580	0.576	**-0.004**
Vartapetiance&Gillam	0.560	0.541	**-0.019**
Grozea	0.480	0.553	**0.073**
Petmanson	0.440	0.671	**0.231**

Combining the Submitted Approaches

- A meta-model can be built based on all the submitted systems
  – A similar idea applied to the PAN-2010 competition on Wikipedia vandalism detection [Potthast et al, 2010]
- A simple meta-classifier is based on the binary output of the 18 submitted models:
  – When the majority of the binary answers is Y/N then a positive/negative answer is produced
  – In ties, a “I don’t know” answer is given
  – A real score is generated, that is the ratio of the number of positive answers to the number of all the answers

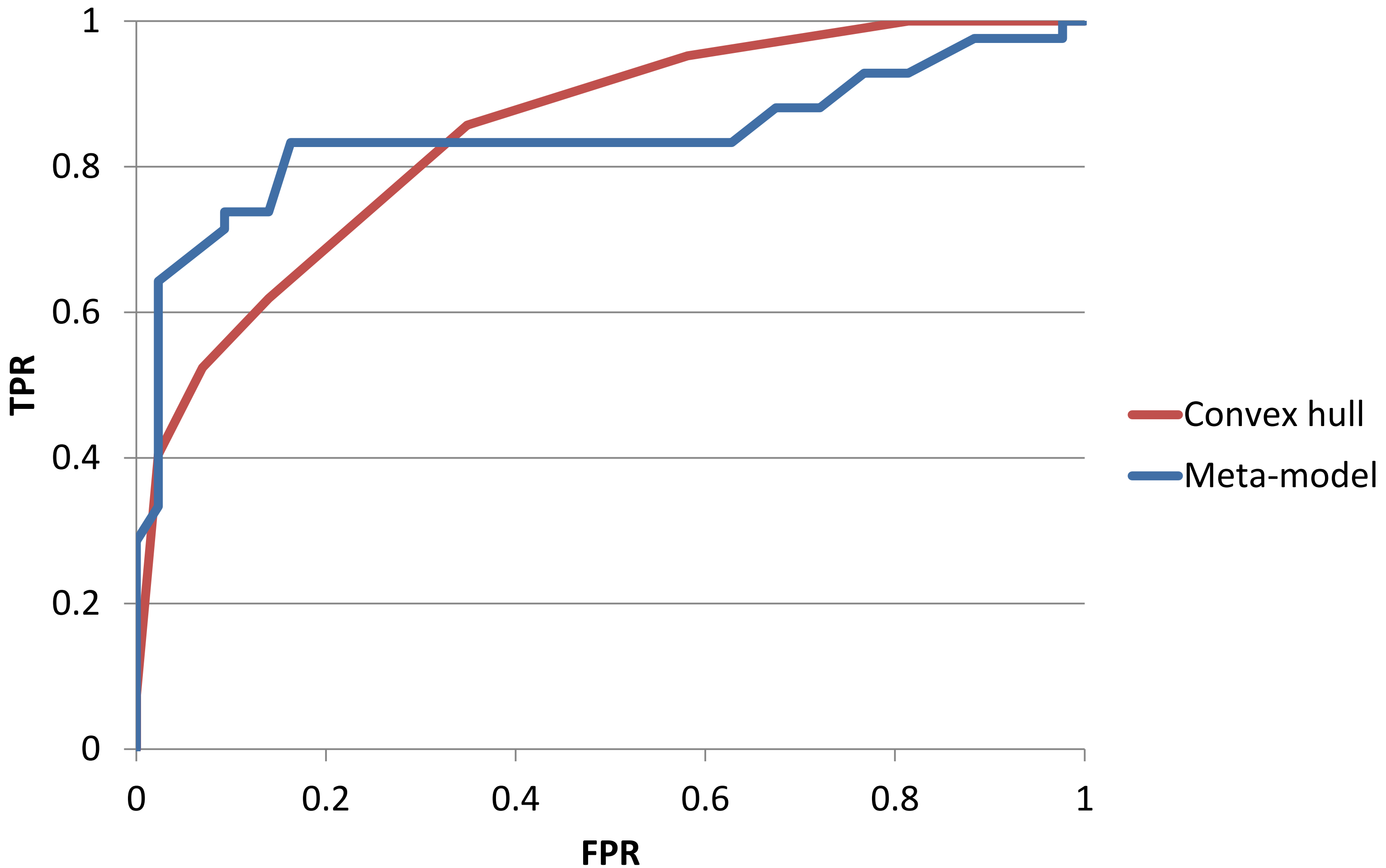
Results of the Meta-model

	F1	Precision	Recall	AUC
Overall	**0.814**	**0.829**	**0.800**	**0.841**
English	**0.867**	**0.867**	**0.867**	0.821
Greek	0.690	0.714	0.667	0.756
Spanish	**0.898**	**0.917**	**0.880**	**0.926**

Seidman's Results:

	F1	Precision	Recall	AUC
Overall	0.753	0.753	0.753	0.735
English	0.800	0.800	0.800	0.792
Greek	**0.830**	**0.830**	**0.830**	**0.824**
Spanish	0.600	0.600	0.600	0.583

Results of the Meta-model (ROC)



Survey of the Submitted Approaches: Text Representation (1)

- Character features
  - letter frequencies, punctuation mark frequencies, character n-grams, common prefixes-suffices, compression-based models
- Lexical features
  - word frequencies, word n-grams, function words, function word n-grams, hapax legomena, morphological information (lemma, stem, case, mood, etc.), word/sentence/paragraph length, grammatical errors and slang words
- Syntactic and semantic features
  - POS n-grams, POS graphs, POS entropy, discourse-level information
  - Considerably increases the computational cost

Survey of the Submitted Approaches: Text Representation (2)

- Combine different types of features in their models
  - [Halvani, *et al.*, Petmanson, *et al.*]
- Use a single type of features
  - [Layton, *et al*., Van Dam]
- Select the most appropriate feature type per language
  - [Seidman]

Survey of the Submitted Approaches: Classification Models (1)

- *Intrinsic* vs. *extrinsic* verification models
- Intrinsic models use only the provided known and unknown documents per problem [Layton *et al*., Halvani *et al*., Jankowska *et al*., Feng&Hirst]
- Extrinsic models use additional external resources (documents from other authors):
  - Taken from the training corpus [Vilariño *et al*.]
  - Downloaded from the Web [Seidman; Veenman&Li]
  - Attempt to transform the one-class classification problem to a binary or multi-class case

Survey of the Submitted Approaches: Classification Models (2)

- Popular classification methods:
  - Ensemble models (very effective in both intrinsic and extrinsic approaches) [Seidman; Halvani, *et al.*; Ghaeini]
  - Modifications of the *CNG* method [Jankowska, *et al.*; Layton et *al.*]
  - Variations of the *unmasking* method [Feng&Hirst; Moreau&Vogel]
  - Compression-based approaches [Bobicev; Veenman&Li]
- The vast majority follow the *instance-based* paradigm
  - Original text-length or equal-size fragments
- Only one approach follows the *profile-based* paradigm [van Dam]

Survey of the Submitted Approaches: Parameter Tuning

- How to optimize the parameter values required by every verification method?
- English/Greek/Spanish:
  - language-dependent parameter settings should be defined
- Some avoid this problem by using global parameter settings [Ghaeini; Halvani, *et al.*; Ledesma, *et al*.]
- The majority estimate the appropriate parameter values per language based on the training corpus
  - Sometimes enhanced by external documents [Jankowska, *et al.*; Petmanson; Seidman]
- Another approach builds an ensemble model using a base classifier for each parameter set configuration [Layton *et al*.]

Survey of the Submitted Approaches: Text Normalization

- The majority did not perform any kind of text preprocessing
  - Use of textual data as found in the given corpus
- Some performed simple transformations
  - Removal of diacritics [van Dam; Halvani, *et al.*]
  - Substitution of digits with a special symbol [van Dam]
  - Conversion of the text to lowercase [van Dam]
- Text-length normalization
  - First concatenate all known documents and then segment them into equal-size fragments [Halvani *et al.*; Bobicev]
  - Reduce all documents within a problem to the same size to produce equal-size representation profiles [Jankowska *et al.*]

Conclusions

- Novelties this year:
  - Focus on a fundamental problem
  - Requirement of software submissions
  - Evaluation corpus covers three languages
- Participation is satisfactory
  - 18 teams from 14 countries
  - Failed attempt to also attract researchers with mainly linguistic background
    - Semi-automated methods

Conclusions

- The most successful approaches follow the extrinsic verification paradigm
- Methods based on complicated NLP-based features do not seem to have any real advantage over simpler methods
  - They also require higher computational cost
- The meta-model combining the output of all the submissions proved to be very effective and in average better than any individual method
  - Heterogeneous models has not attracted much attention so far in authorship attribution research

Conclusions

- The vast majority of the participants answered all the problems
  - This makes Precision and Recall measures equal
  - Only two teams used the “I don’t know” option
- Better evaluation criteria are needed focusing on the ability of the models to only provide quasi-certain answers
  - E.g., c@1 used in the question answering community
  - Mandatory use of real scores indicating the confidence of the provided answers

Thank you for your participation!

Your suggestions for improving future PANs are particularly welcome!