# ECML/PKDD 2012 Discovery Challenge: *PASCAL Large Scale Hierarchical Text Classification*

I. Partalas*, A. Kosmopoulos†,◇, G. Paliouras†, E. Gaussier*,
I. Androutsopoulos◇, T. Artières‡, P. Gallinari‡

* Lab. d'Informatique de Grenoble & Grenoble University, France
† National Center for Scientific Research "Demokritos", Greece
◇ Athens University of Economics and Business, Greece
‡ Lab. d'informatique de Paris 6, France

May 13, 2014

# Large scale hierarchical classification (1)

- Large volumes of data (instances, features, classes)

Examples:

- DMOZ: over 600,000 classes
- Wikipedia: over 700,000 classes

# Large scale hierarchical classification (1)

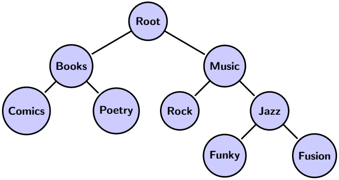
- Large volumes of data (instances, features, classes)
- Research efforts strive to address large-scale problems [Xue et al., 2008],[S. Bengio and Grangier, 2010], [Zhao et al., 2011],

**Challenges on Large-scale Learning:**

- Large-scale Hierarchical Text Classification
- Imagenet Large Scale Visual Recognition

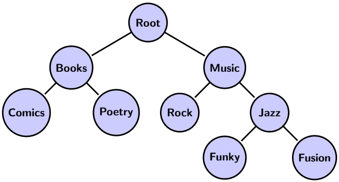
# Large scale hierarchical classification (1)

- Large volumes of data (instances, features, classes)
- Research efforts strive to address large-scale problems [Xue et al., 2008],[S. Bengio and Grangier, 2010], [Zhao et al., 2011],
- Exploitation of semantic relations among the classes



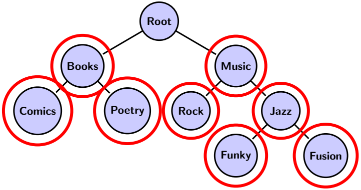
# Large scale hierarchical classification (2)

- Hierarchical
  - Top-down approaches (per class, per parent, per level)



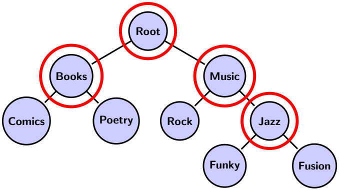
# Large scale hierarchical classification (2)

- Hierarchical
  - Top-down approaches (per class, per parent, per level)



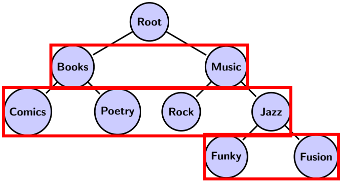
# Large scale hierarchical classification (2)

- Hierarchical
  - Top-down approaches (per class, per parent, per level)



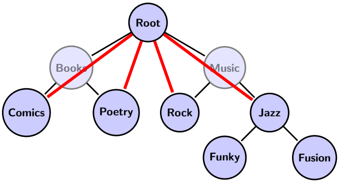
# Large scale hierarchical classification (2)

- Hierarchical
  - Top-down approaches (per class, per parent, per level)



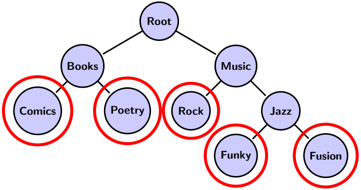
# Large scale hierarchical classification (2)

- Hierarchical
  - Top-down approaches (per class, per parent, per level)
- Mildly hierarchical
  - Usually a sub-part of the hierarchy is used (flattened)



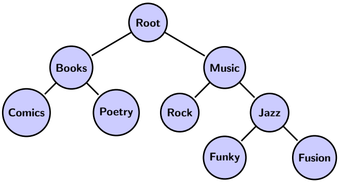
# Large scale hierarchical classification (2)

- Hierarchical
  - Top-down approaches (per class, per parent, per level)
- Mildly hierarchical
  - Usually a sub-part of the hierarchy is used (flattened)
- Flat



# Large scale hierarchical classification (2)

- Hierarchical
  - Top-down approaches (per class, per parent, per level)
- Mildly hierarchical
  - Usually a sub-part of the hierarchy is used (flattened)
- Flat




**Challenges:**

- LSHTC3 best system: 38% Acc. (Large Wikipedia, 325K classes)
- Scale to even more classes
- Take into account the complex relationships among the classes

# Past Challenges

**LSHTC1:**

- Data source: ODP Web directory
- Tracks: Basic, Cheap, Expensive, Full
- Hierarchy type: tree
- Max num of categories: 12,000

**LSHTC2:**

- Data source: ODP Web directory and Wikipedia
- Tracks: DMOZ (27K), Wikipedia small (36K), Wikipedia large (325K)
- Hierarchy type: tree and DAG
- Max num of categories: 325,000
- Multi-label data

# LSHTC3

- Track 1: Large Scale Hierarchical Classification
  - Wikipedia dataset
  - Task 1: Medium-size (36,500 classes)
  - Task 2: Large (325,000 classes)
- Track 2: Multi-task Learning
  - DMOZ and Wikipedia medium size
  - 12,000 classes each
- Track 3: Refinement Learning
  - DMOZ dataset
  - Task 1: semi-supervised
  - Task 2: unsupervised

# Track 1: Large Scale Hierarchical Classification

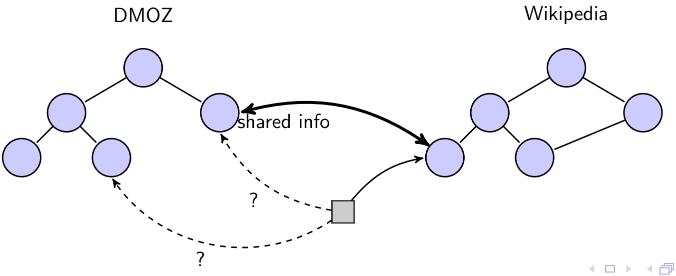
- 2 versions of Wikipedia dataset
- Task 1: medium-size (36,500 classes)
  - Original text data
  - Pre-processed data
- Task 2: large Wikipedia (325,000 classes)
- Multi-label and hierarchy is DAG

# Track 2: Multi-task learning

- Wikipedia and DMOZ datasets
- Common feature space
- 12,000 classes for each dataset
- Single-label, hierarchy: DAG for Wikipedia and tree for DMOZ

**Goal:**

Use shared information in order to improve performance on each task

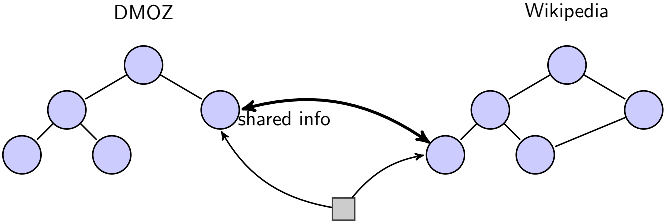


# Track 2: Multi-task learning

- Wikipedia and DMOZ datasets
- Common feature space
- 12,000 classes for each dataset
- Single-label, hierarchy: DAG for Wikipedia and tree for DMOZ

**Goal:**

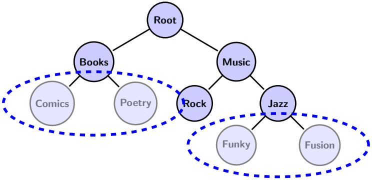
Use shared information in order to improve performance on each task



# Track 3: Refinement Learning

## Task 1: semi-supervised

- A reduced (12,000 classes) and an expanded (14,000 classes) hierarchy is available
- Two documents are given for each expanded class
- Goal: to reassign test documents to the new classes



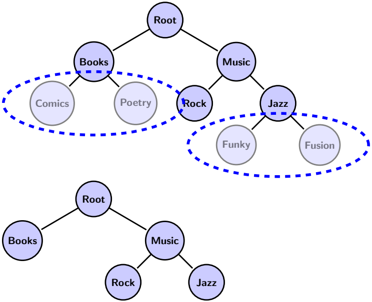
# Track 3: Refinement Learning

**Task 1: semi-supervised**

- A reduced (12,000 classes) and an expanded (14,000 classes) hierarchy is available
- Two documents are given for each expanded class
- Goal: to reassign test documents to the new classes

**Task 2: unsupervised**

- Only the reduced hierarchy is given
- Goal: expand the hierarchy



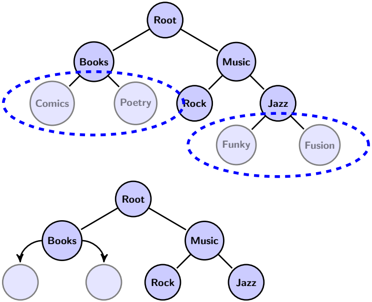
# Track 3: Refinement Learning

## Task 1: semi-supervised

- A reduced (12,000 classes) and an expanded (14,000 classes) hierarchy is available
- Two documents are given for each expanded class
- Goal: to reassign test documents to the new classes

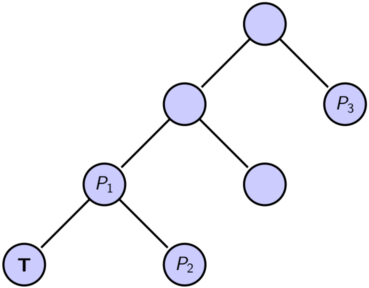
## Task 2: unsupervised

- Only the reduced hierarchy is given
- Goal: expand the hierarchy



# Flat and Hierarchical measures

- T is the correct category
- $P_1, P_2, P_3$ are the predicted categories
- Flat measures treat the errors of $P_1$, $P_2$ and $P_3$ in the same way
- A hierarchical measure should penalize differently each error



# Multi-label - Example based [Tsoumakas et al., 2010]

$$Accuracy = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}$$

$$F_1 = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|}$$

- $D$ is the number of testing documents
- $Z_i$ the labels predicted by the classifier
- $Y_i$ the true labels of the document

# Multi-label - Label based [Tsoumakas et al., 2010]

$$M_{macro} = \frac{1}{|L|} \sum_{\lambda=1}^{|L|} M(tp_\lambda, fp_\lambda, tn_\lambda, fn_\lambda)$$

$$M_{micro} = M(\frac{1}{|L|} \sum_{\lambda=1}^{|L|} tp_\lambda, \frac{1}{|L|} \sum_{\lambda=1}^{|L|} fp_\lambda, \frac{1}{|L|} \sum_{\lambda=1}^{|L|} tn_\lambda, \frac{1}{|L|} \sum_{\lambda=1}^{|L|} fn_\lambda)$$

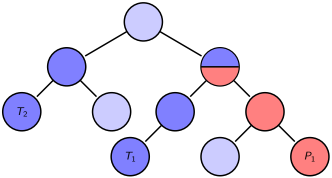
where $L$ represents the labels and $M$ can be either precision or recall

# Hierarchical versions of $P$ and $R$, $F_1$ [Costa et al., 2007]

$$HP = \frac{|An(C_p) \cap An(C_t)|}{|An(C_p)|}$$

$$HR = \frac{|An(C_p) \cap An(C_t)|}{|An(C_t)|}$$

- $C_p$ is the set of predicted categories
- $An(C_p)$ is the set of ancestors of $C_p$
- $C_t$ is the set of true categories
- $An(C_t)$ is the set of ancestors of $C_t$



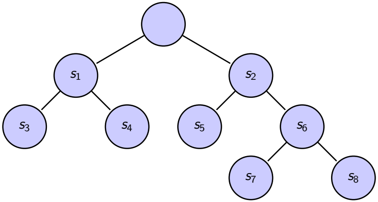

HP = 1/3, HR=1/5

# Participating Systems

- Number of participants:
  - Track 1 - medium: 16
  - Track 1 - large: 5
  - Track 2: 3
  - Track 3 - semi-supervised: 0
  - Track 3 - unsupervised: 1
- Total submissions: 900

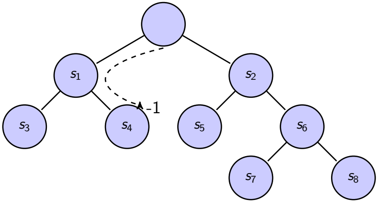
# Overview of Approaches (I) - Supervised

- Arthur [Wang et al., 2011]
  - Meta-learning problem based on hierarchical TD classification
  - Meta-features: scores of base-classifiers towards the leaves
  - Meta-label: +1 for correct classification, -1 otherwise



# Overview of Approaches (I) - Supervised

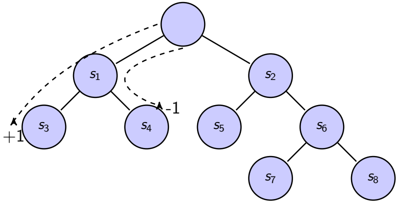
- Arthur [Wang et al., 2011]
  - Meta-learning problem based on hierarchical TD classification
  - Meta-features: scores of base-classifiers towards the leaves
  - Meta-label: +1 for correct classification, -1 otherwise




Inst. 1 1:$s_1$ 4:$s_4$ -1

# Overview of Approaches (I) - Supervised

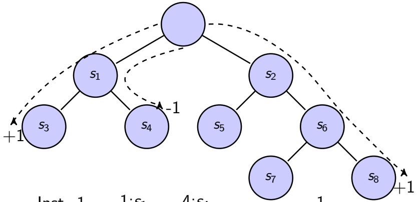
- Arthur [Wang et al., 2011]
  - Meta-learning problem based on hierarchical TD classification
  - Meta-features: scores of base-classifiers towards the leaves
  - Meta-label: +1 for correct classification, -1 otherwise




| | | | |
|---|---|---|---|
| Inst. 1 | 1:$s_1$ | 4:$s_4$ | -1 |
| Inst. 2 | 1:$s_1$ | 3:$s_3$ | +1 |

# Overview of Approaches (I) - Supervised

- Arthur [Wang et al., 2011]
  - Meta-learning problem based on hierarchical TD classification
  - Meta-features: scores of base-classifiers towards the leaves
  - Meta-label: +1 for correct classification, -1 otherwise




| | | | | |
|---|---|---|---|---|
| Inst. 1 | 1:$s_1$ | 4:$s_4$ | | -1 |
| Inst. 2 | 1:$s_1$ | 3:$s_3$ | | +1 |
| Inst. 3 | 2:$s_2$ | 5:$s_5$ | 8:$s_8$ | +1 |

# Overview of Approaches (I) - Supervised

- Arthur [Wang et al., 2011]
  - Meta-learning problem based on hierarchical TD classification
  - Meta-features: scores of base-classifiers towards the leaves
  - Meta-label: $+1$ for correct classification, -1 otherwise
- TTI [Sasaki and Weissenbacher, 2012]
  - Top-down scheme (a classifier for each parent-child)
  - Thresholding adjustment using the scores of the SVMs
  - Pruning of the final labels below threshold

# Overview of Approaches (II) - Supervised

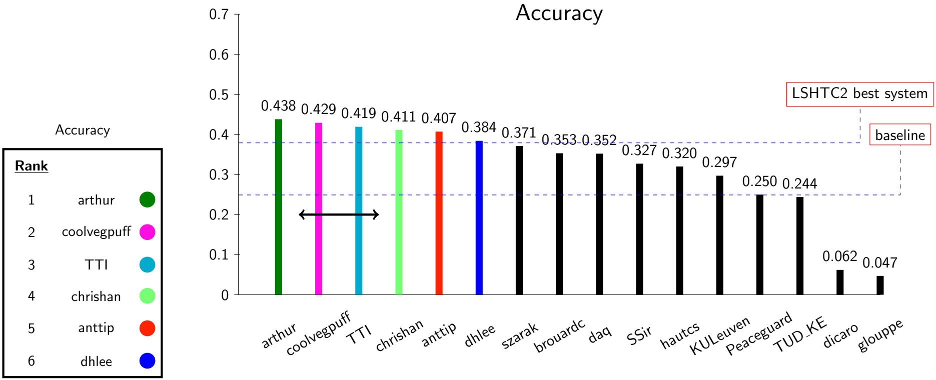
- Anttip [Puurula and Bifet, 2012]
  - Flat classification
  - Ensemble of optimized MNB
  - A greedy pruning algorithm is adopted
- Chrishan [Han et al., 2012]
  - k-NN based
  - Combines two similarity measures
  - Hierarchical information is incorporated in the ranking procedure
- Dhlee [Lee, 2012]
  - Flat approach
  - Based on Rocchio classification
  - Uses Label-Power set transformation for multi-labeling
  - Limits the predicted label set with a greedy search

# Overview of Approaches (III) - Unsupervised

Marcacini [Marcacini et al., 2012]

- 3 basic steps
- 1: a category is selected for expansion
- 2: a hierarchical clustering algorithm is applied and a dendogram is derived
- 3: the new categories are refined

# Track 1 - Wikipedia medium (I)




⟷ No significant statistical difference [Yang and Liu, 1999]

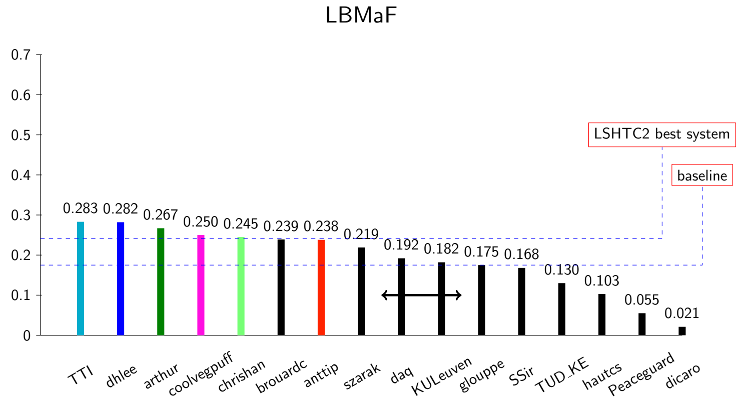
# Track 1 - Wikipedia medium (I)

Accuracy

| Rank | | |
|---|---|---|
| 1 | arthur | |
| 2 | coolvegpuff | |
| 3 | TTI | |
| 4 | chrishan | |
| 5 | anttip | |
| 6 | dhlee | |

LBMaF

| Rank | | |
|---|---|---|
| 1 | TTI | |
| 2 | dhlee | |
| 3 | arthur | |
| 4 | coolvegpuff | |
| 5 | chrishan | |
| 6 | brouardc | |
| 7 | anttip | |

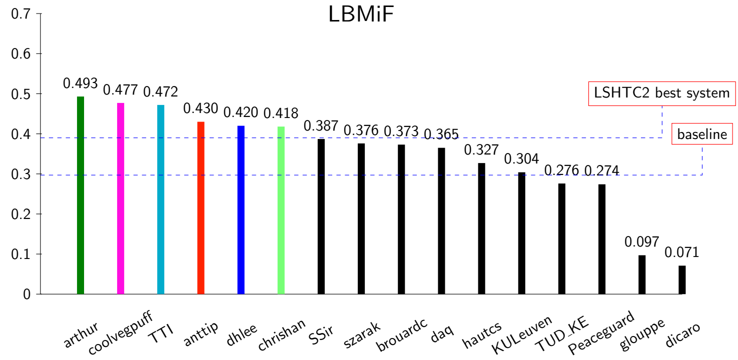



⟷ No significant statistical difference [Yang and Liu, 1999]

# Track 1 - Wikipedia medium (II)

LBMiF

| Rank | | |
|---|---|---|
| 1 | arthur | |
| 2 | coolvegpuff | |
| 3 | TTI | |
| 4 | anttip | |
| 5 | dhlee | |
| 6 | chrishan | |



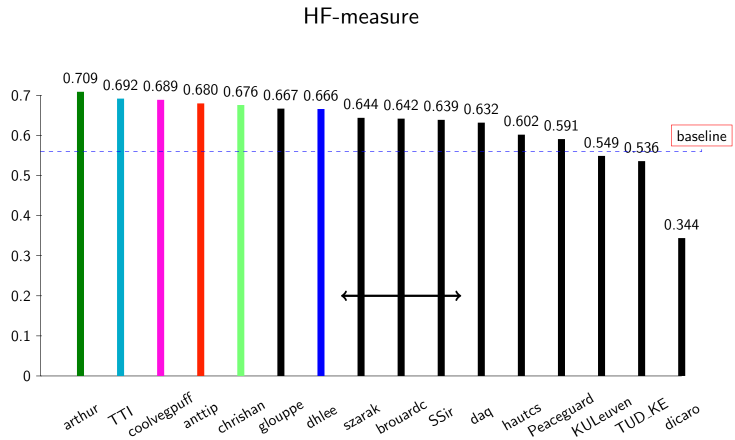
# Track 1 - Wikipedia medium (II)

LBMiF

| Rank | |
|---|---|
| 1 | arthur |
| 2 | coolvegpuff |
| 3 | TTI |
| 4 | anttip |
| 5 | dhlee |
| 6 | chrishan |

HF

| Rank | |
|---|---|
| 1 | arthur |
| 2 | TTI |
| 3 | coolvegpuff |
| 4 | anttip |
| 5 | chrishan |
| 6 | glouppe |
| 7 | dhlee |




⟷ No significant statistical difference [Yang and Liu, 1999]

# Wikipedia medium - Summary

**Accuracy**

| Rank | |
|---|---|
| 1 | arthur |
| 2 | coolvegpuff |
| 3 | TTI |
| 4 | chrishan |
| 5 | anttip |
| 6 | dhlee |
| 7 | szarak |

**LBMaF**

| Rank | |
|---|---|
| 1 | TTI |
| 2 | dhlee |
| 3 | arthur |
| 4 | coolvegpuff |
| 5 | chrishan |
| 6 | brouardc |
| 7 | anttip |

**LBMiF**

| Rank | |
|---|---|
| 1 | arthur |
| 2 | coolvegpuff |
| 3 | TTI |
| 4 | anttip |
| 5 | dhlee |
| 6 | chrishan |
| 7 | SSir |

**HF**

| Rank | |
|---|---|
| 1 | arthur |
| 2 | TTI |
| 3 | coolvegpuff |
| 4 | anttip |
| 5 | chrishan |
| 6 | glouppe |
| 7 | dhlee |

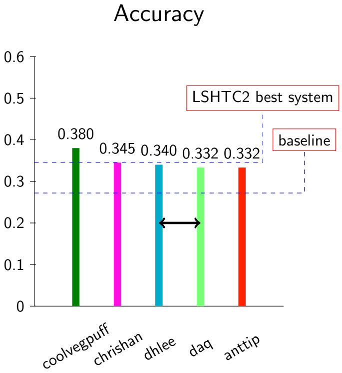
## Observations

- The three best systems in first places across most the measures
- TTI and dhlee perform better in rare categories (best LBMaF scores)
- glouppe is ranked 6th in HF measure (average number of labels 10.64, predicts internal nodes mostly (90%))

# Track 1 - Wikipedia Large (I)

Accuracy

| **Rank** | | |
|---|---|---|
| 1 | coolvegpuff | |
| 2 | chrishan | |
| 3 | dhlee | |
| 4 | daq | |
| 5 | anttip | |



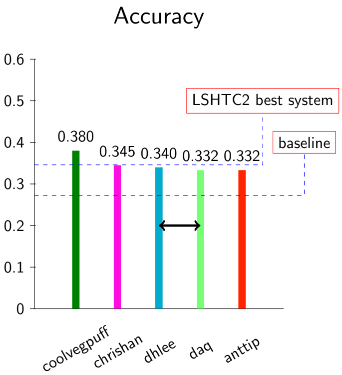
# Track 1 - Wikipedia Large (I)

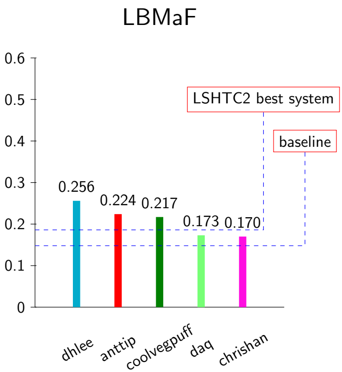
Accuracy

| **Rank** | | |
|---|---|---|
| 1 | coolvegpuff | 🟢 |
| 2 | chrishan | |
| 3 | dhlee | |
| 4 | daq | |
| 5 | anttip | |

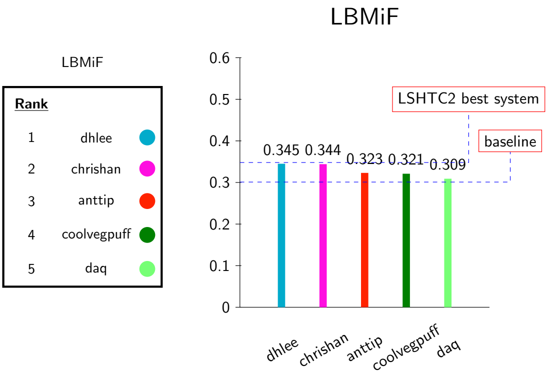
LBMaF

| **Rank** | | |
|---|---|---|
| 1 | dhlee | |
| 2 | anttip | |
| 3 | coolvegpuff | |
| 4 | daq | |
| 5 | chrishan | |



# Track 1 - Wikipedia Large (II)



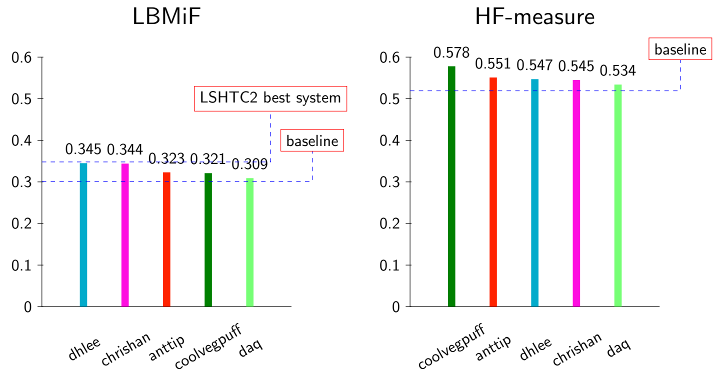
# Track 1 - Wikipedia Large (II)

LBMiF

| Rank | | |
|---|---|---|
| 1 | dhlee | ● |
| 2 | chrishan | ● |
| 3 | anttip | ● |
| 4 | coolvegpuff | ● |
| 5 | daq | ● |

HF

| Rank | | |
|---|---|---|
| 1 | coolvegpuff | ● |
| 2 | anttip | ● |
| 3 | dhlee | ● |
| 4 | chrishan | ● |
| 5 | daq | ● |



# Track 1 - Wikipedia Large Summary

| Rank | Accuracy | LBMaF | LBMiF | HF |
|---|---|---|---|---|
| 1 | coolvegpuff | dhlee | dhlee | coolvegpuff |
| 2 | chrishan | anttip | chrishan | anttip |
| 3 | dhlee | coolvegpuff | anttip | dhlee |
| 4 | daq | daq | coolvegpuff | chrishan |
| 5 | anttip | chrishan | daq | daq |

## Observations

- Differences in the systems (different behavior across the measures)
- Dhlee balances precision and recall

# Track 1 - Wikipedia Large Summary

Accuracy

| Rank | |
|---|---|
| 1 | coolvegpuff |
| 2 | chrishan |
| 3 | dhlee |
| 4 | daq |
| 5 | anttip |

LBMaF

| Rank | |
|---|---|
| 1 | dhlee |
| 2 | anttip |
| 3 | coolvegpuff |
| 4 | daq |
| 5 | chrishan |

LBMiF

| Rank | |
|---|---|
| 1 | dhlee |
| 2 | chrishan |
| 3 | anttip |
| 4 | coolvegpuff |
| 5 | daq |

HF

| Rank | |
|---|---|
| 1 | coolvegpuff |
| 2 | anttip |
| 3 | dhlee |
| 4 | chrishan |
| 5 | daq |

## Observations

- Differences in the systems (different behavior across the measures)
- Dhlee balances precision and recall
- F-measure problem: For two systems A and B, if A.precision>>B.precision and A.recall<B.recall then is possible for A.f-measure < B.f-measure

# Track 1 - Wikipedia Large Summary

**Accuracy**

| **Rank** | | |
|---|---|---|
| 1 | coolvegpuff | ● |
| 2 | chrishan | ● |
| 3 | dhlee | ● |
| 4 | daq | ● |
| 5 | anttip | ● |

**LBMaF**

| **Rank** | | |
|---|---|---|
| 1 | dhlee | ● |
| 2 | anttip | ● |
| 3 | coolvegpuff | ● |
| 4 | daq | ● |
| 5 | chrishan | ● |

**LBMiF**

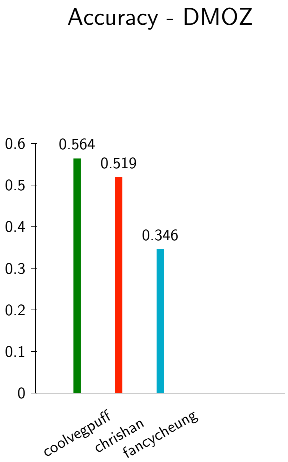
| **Rank** | | |
|---|---|---|
| 1 | dhlee | ● |
| 2 | chrishan | ● |
| 3 | anttip | ● |
| 4 | coolvegpuff | ● |
| 5 | daq | ● |

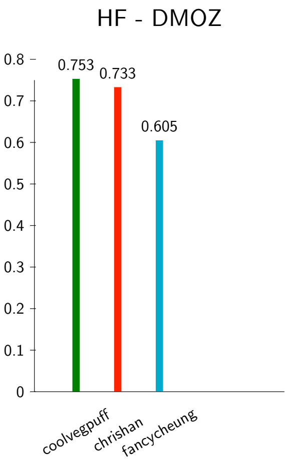
**HF**

| **Rank** | | |
|---|---|---|
| 1 | coolvegpuff | ● |
| 2 | anttip | ● |
| 3 | dhlee | ● |
| 4 | chrishan | ● |
| 5 | daq | ● |

| | LBMiP | LBMiR | LBMiF |
|---|---|---|---|
| chrishan | 0.551 | 0.250 | 0.344 b |
| dhlee | 0.415 | 0.295 | 0.345 |

# Track 2 - Multi-task Learning



# Track 3 - Unsupervised

- Marcacini system
- HF-measure: 0.354
- Precision: 0.841
- Recall: 0.285

# Conclusions

- A variety of hierarchical and flat approaches
- Participants focused in Track 1
- 5 participants from LSHTC2 participated to the new challenge
- Better results in same tracks of LSHTC2 (14% for medium, 8% for large)
- We are not aware if any pre-processing steps were used in Wikipedia medium

# Conclusions

- A variety of hierarchical and flat approaches
- Participants focused in Track 1
- 5 participants from LSHTC2 participated to the new challenge
- Better results in same tracks of LSHTC2 (14% for medium, 8% for large)
- We are not aware if any pre-processing steps were used in Wikipedia medium

**Next challenges**

LSHTC4 and BioASQ challenges

# Questions

Thank you for your attention!

# Open Issues

## Evaluation Measures

- Clear differences among flat and hierarchical measures
- Do these measures suffice for evaluation?

## How to attract researchers in Tracks 2 and 3?

- Was there something that prevented researchers to participate into these tracks?
- Why would you participate in such tracks?

## How it can be related to other challenges?

- Large Scale Visual Recognition Challenge (http://www.image-net.org/)
- Creation of a common challenge?
- Wide use as benchmark.

# BioASQ Challenge

- Challenge on biomedical semantic indexing and Question-Answering
- Motivating example: *Q1: What is the role of thyroid hormones administration in the treatment of heart failure?*

## Objectives

1. Large-scale classification of biomedical documents onto ontology concepts, in order to automate semantic indexing
2. classification of biomedical questions onto the same concepts
3. integration of relevant document snippets, information databases and knowledge bases, and
4. delivery of the retrieved information in a concise and user-understandable form

Workshops will be organized dedicated to the challenge

# The Challenge

- Participant: BioMedAnswers Inc.

**Task 1a**

- BioASQ distributes new unclassified PubMeed abstracts
- BioMedAnswers attaches MeSH terms (limited resp. time)
- Evaluation when abstracts get classified by PubMed curators

**Task 1b**

Stage A

- BioASQ distributes questions from benchmark
- BioMedAnswers responds with concepts, snippets, triples

Stage B

- BioASQ distributes questions + concepts, snippets, triples
- BioMedAnswers responds with facts, summaries, etc.

Evaluation with gold answers, majority and manually (sample)

# The Challenge (II)

**Task 2a**

- Same as 1a, with new data and improvements

**Task 2b**

Similar to 1b

- BioASQ distributes questions from new benchmark
- BioMedAnswers responds with concepts, snippets, triples, facts summaries, etc.

Evaluation with gold answers, majority and manually (sample)
Each type of response evaluated separately

# Significance Tests - for Macro measures

Macro sign test (S-test)Yang and Liu [1999]

$$Z = \frac{k - 0.5n}{0.5\sqrt{n}}, \text{since } n > 12$$

# Significance Tests - for Micro measures, $HF_1$, HP and HR

Micro sign test (S-test)Yang and Liu [1999]

$$Z = \frac{k - 0.5n}{0.5\sqrt{n}}, \text{since } n > 12$$

- $n$ is the number of times that $a_i$ and $b_i$ differ
- $k$ is he number of times that $a_i$ is larger than $b_i$
- $a_i \in \{0, 1\}$ is the measure of success for system $A$ on the $i$th decision (i= 1, 2, ..., N)
- $b_i \in \{0, 1\}$ is the measure of success for system $B$ on the $i$th decision (i= 1, 2, ..., N)
- $N$ is the number of binary decisions
- Significant different if P-value $< 0.05$

# Significance Tests

- The null hypothesis is that $k$ has a binomial distribution Bin(n, p) where $p = 0.5$
  $\Rightarrow$ there is no significant difference between the two systems
- The alternative hypothesis is that he binomial distribution of $k$ with $p > 0.5$
  $\Rightarrow$ system A is better than system B
- A larger difference doesn't always translate to significant difference
- Abnormality in significant difference between systems ranked by an evaluation measure
  For example:
  - $A > B > C$ according to evaluation measure X
  - But A appears significantly better than B but not than C

E.P. Costa, A.C. Lorena, A.C.P.L.F. Carvalho, and A.A. Freitas. A review of performance evaluation measures for hierarchical classifiers. In *Evaluation Methods for Machine Learning II: papers from the AAAI-2007 Workshop, AAAI Technical Report WS-07-05*, pages 1–6, July 2007.

Xiaogang Han, Shaohua Li, and Zhiqi Shen. A k-nn method for large scale hierarchical text classification at lshtc3. In *ECML/PKDD Discovery Challenge Workshop: Pascal Large Scale Hierarchical Classification*, 2012.

Dong-Hyun Lee. Multi-stage rocchio classification for large-scale multi-labeled text data. In *ECML/PKDD Discovery Challenge Workshop: Pascal Large Scale Hierarchical Classification*, 2012.

Ricardo Marcondes Marcacini, Everton A. Cherman, Jean Metz, and Solange O. Rezende. A fast dendrogram refinement approach for unsupervised expansion of hierarchies. In *ECML/PKDD Discovery Challenge Workshop: Pascal Large Scale Hierarchical Classification*, 2012.

Antti Puurula and Albert Bifet. Ensembles of sparse multinomial classi
ers for scalable text classi
cation. In *ECML/PKDD Discovery Challenge Workshop: Pascal Large Scale Hierarchical Classification*, 2012.

J. Weston S. Bengio and D. Grangier. Label embedding trees for large multi-class tasks. In *NIPS*, 2010.

Yutuka Sasaki and Davy Weissenbacher. Tti's system for the lshtc3 challenge. In *ECML/PKDD Discovery Challenge Workshop: Pascal Large Scale Hierarchical Classification*, 2012.

G. Tsoumakas, I. Katakis, and I. Vlahavas. Random k-labelsets for multi-label classification. In *IEEE Transactions on Knowledge Discovery and Data Engineering*, 2010.

Xiao-Lin Wang, Hai Zhao, and Bao-Liang Lu. Enhance top-down method with meta-classification for very large-scale hierarchical classification. In *International Joint Conference on Natural Language Processing*, pages 1089–1097, 2011.

Gui-Rong Xue, Dikan Xing, Qiang Yang, and Yong Yu. Deep classification in large-scale text hierarchies. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 619–626, 2008.

Yiming Yang and Xin Liu. A re-examination of text categorization methods. pages 42–49. ACM Press, 1999.

Bin Zhao, Fei Fei F. Li, and Eric P. Xing. Large-scale category structure aware image categorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1251–1259, 2011.