# Beyond Deep Learning: Enriching Data Representations for Machine Learning Tasks

Nikiforos Pittaras

Ph.D. Defense

Department of Informatics & Telecommunications
National and Kapodistrian University of Athens

Institute of Informatics & Telecommunications
NCSR "Demokritos"

*npittaras@di.uoa.gr*

August 31, 2021

## Acknowledgements


- Ph.D. proposal awared by a Industrial Ph.D. Scholarship Program of the Stavros Niarchos Foundation.
- Industrial application component

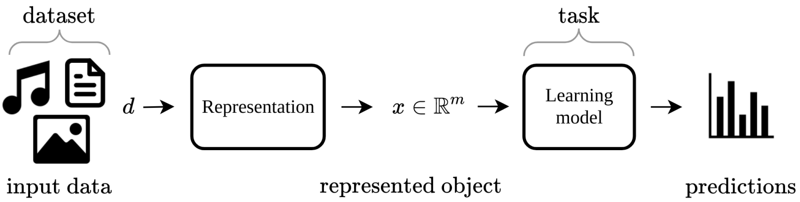
# Introduction

# Motivation

Focus: Machine learning (ML) systems

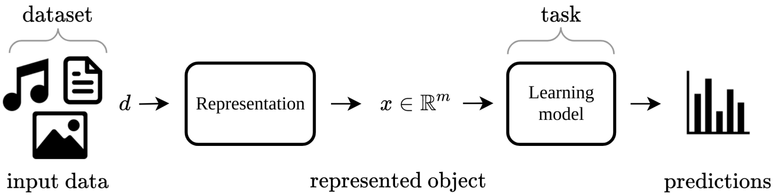
- Applications crucial for social / scientific / commercial ecosystems
- E.g. classification, clustering, summarization solutions
- Improving such systems can yield significant benefits

# Typical Machine Learning Pipeline




- Dataset: real-world objects / ground truth $d$
- Representation: Maps $d$ to a vector format
- Represented objects: Vector format $x$
- Learning model: finds associations / patterns in $x$
- Predictions: useful information produced by learning model

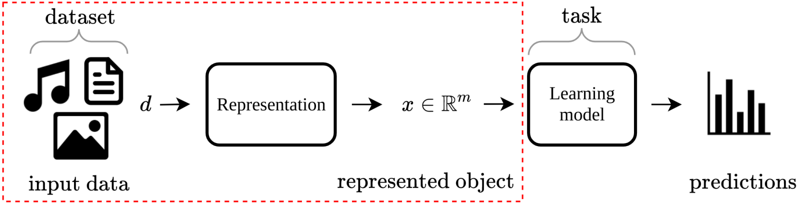
# Improving ML pipeline performance




Indicative avenues for ML system improvement

- Resource-oriented:
  - More compute (GPUs / training times)
  - Greater quantity / quality of training data
- Modelling-oriented:
  - Improve the representation approach
  - Improve the learning model

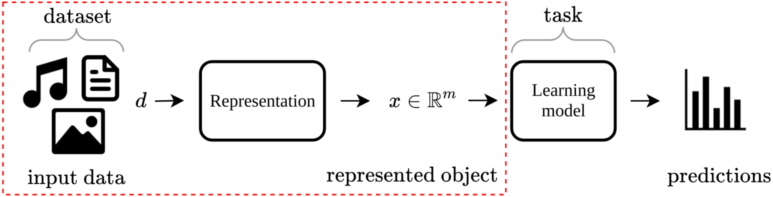
# Improving ML pipeline performance




Indicative avenues for ML system improvement

- Resource-oriented:
  - More compute (GPUs / training times)
  - Greater quantity / quality of training data
- Modelling-oriented:
  - **Improved representation approach**
  - Improve the learning model

# Importance of Representations




- Early step in the pipeline, benefits / errors propagate
- $x$ : abstraction of $d$: May discard noise / lose information
- Important semantics/context may/may not be included in $x$
- Only input to Learning model: $x$
- Thus, *semantic gap* between $x$ and $d$ impacts performance

# Representation Enrichment

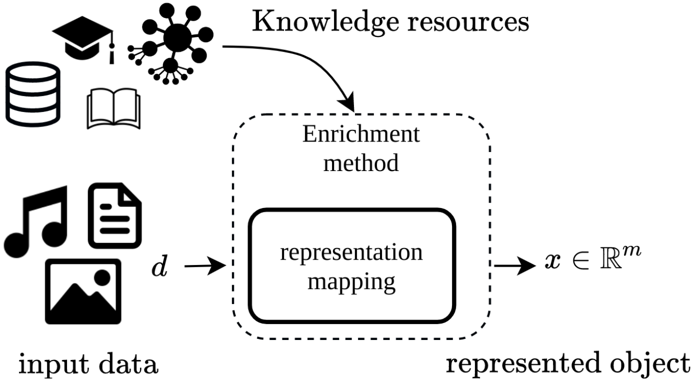
How can we narrow / bridge the semantic gap?

- Utilize resources of curated, high-level, structured knowledge (e.g. ontologies, lexicons, knowledge bases, class hierarchies)
- **Go beyond content-based representations via enrichment**

# Representation Enrichment

How can we narrow / bridge the semantic gap?

- Utilize resources of curated, high-level, structured knowledge (e.g. ontologies, lexicons, knowledge bases, class hierarchies)
- **Go beyond content-based representations via enrichment**



# Goals

Holistic study of representations for ML problems, including:

- Content-based representation approaches
- Methods for knowledge-based representation enrichment
- Available structured human knowledge resources

Broad investigation:

- Different ML tasks (classification, clustering, summarization)
- Different data modalities (text, images, audio)

# Contributions

Content-based Representations:

- Literature review [11]
- Novel proposals / applications [2][3][4][5][12][13]

Representation Enrichment:

- Overview of different exploitable knowledge resources [11]
- Literature review for representation enrichment methods [11]
- Novel proposals for enriching different ML tasks [1][10][11]
- Consolidation of findings to an industrial ML application [16]

# Content-based Methods

# Focus of the Study

Study scope [1]:

- Content-based: consider only intra-instance content
- No additional/external information sources
- Focus on the context of classification
- Text, image, audio data modalities

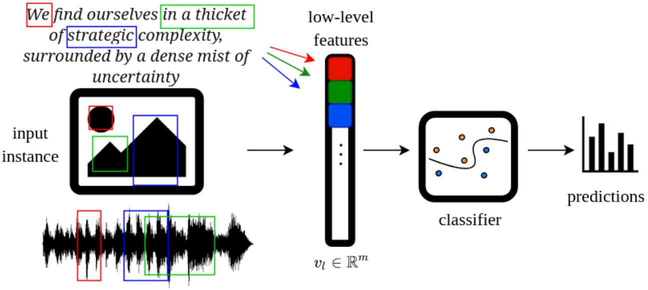
Contributions: identified three broad paradigms

1. Low-level / template-matching representations
2. Aggregation-based representations
3. Deep representation learning approaches

[1] Pittaras et al., Content-based and knowledge-enriched representations for classification across modalities: a survey, ACM CSUR (under review)

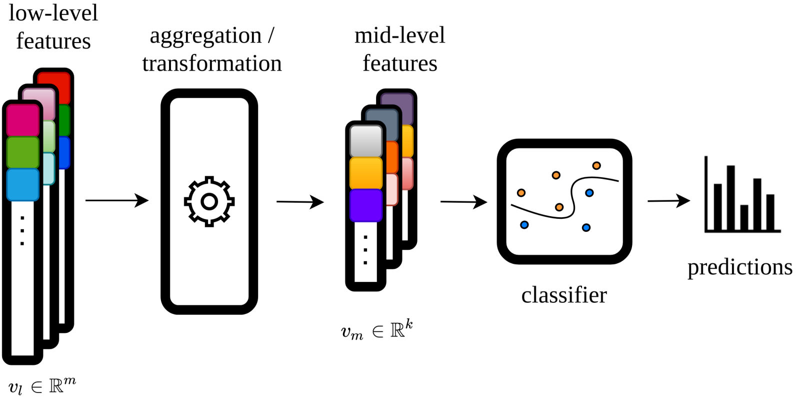
# Low-level / Template-Matching

- Locally / globally apply preconfigured templates
- Template output responses used as features
- E.g. Bag of Words / Features, simple input statistics, visual / audio descriptors



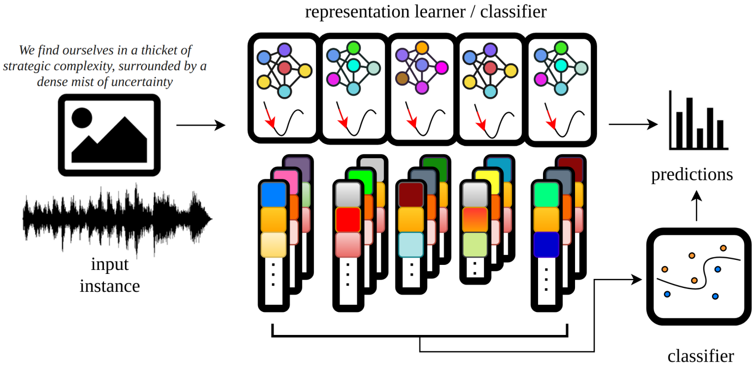
# Aggregation-based

- Utilize ensembles of low-level representation instances
- Improve by applying pre-defined, engineered processing steps
- Transform / combine into (reduced) distributed latent space
- E.g. Clustering, factorization/decomposition, topic modelling



# Deep representation learning

- Non-linear hierarchies (simple to rich), distributed features
- End-to-end task & representation learning
- High pretraining & transfer-learning capabilities
- E.g. Feedforward, convolutional, recurrent NNs



# Per-paradigm Pros/Cons

Generally observed:

| paradigm / attributes | low-level | aggregation | deep |
|---|---|---|---|
| high-level semantics | X | ? | ✓ |
| explainable | ✓ | ? | X |
| data-driven / learned | X | ? | ✓ |
| low-dimensional / space-efficient | ? | ✓ | ✓ |
| data efficient / lean | ✓ | ✓ | X |
| computationally efficient | ? | X | X |

# Findings

- Multiple representation paradigms
- Strength and weaknesses for each; no one-size-fits-all
- Paradigm evolution: low $\rightarrow$ aggregation-based $\rightarrow$ deep
- Evolution reflects search for rich, informative features

# Overview

Motivation:

- Improve understanding accross different applications
- Identify task / domain-specific challenges / points of improvement

Broad investigation:

- Across tasks (classification, summarization, clustering)
- In conjunction with different, diverse learning models
- Across different domains / modalities

# Focused studies

- Hate Speech Detection [3]
  - Text, Classification, Deep Word Embeddings, NGGs
- Extractive Summarization of Web Documents [2]
  - Text, Summarization, Topic-based features
- Automatic Summarization of Video Game Reviews [5]
  - Text, Summarization, Novel domain, Deep Embeddings
- Documents / Social Media Analysis in the Security Domain [4]
  - Text, Clustering, Classification, Summarization, NGGs
- Scaling and Enrichment of Automatic Summarization [13]
  - Text, Summarization, Performance Scaling, Utilization of NER information
- Classifying Videos with Multimodal DNNs [12]
  - Video (Image, Audio), Classification, Deep Features

# Summary of Content-based Representations

- Multiple, diverse approaches; no one-size-fits-all method
- Indications for no-free-lunch theorem for representations
- Trend towards semantically rich representations
- Richness beneficial to multiple tasks, domains and modalities

→ **Strengthen motivation to examine representation enrichment**

# Representation Enrichment Approaches

# Representation Enrichment

Thesis focus:

- Representation enrichment with human knowledge can improve task performance

Enriching with human knowledge may address:

- Missing contextual information
- Missing domain-specific knowledge
- Ambiguity in the data and their generation process
- Need for transparency & explainability

# Focus of the Study

Study [1] scope:

- Enrichment: look beyond instance content
- Mine resources of structured knowledge
- Focus on the context of classification
- Text, image, audio data modalities

Contributions:

- Summary of structured knowledge resources
- Identified three enrichment paradigms

[1] Pittaras et al., Content-based and knowledge-enriched representations for classification across modalities: a survey, ACM CSUR (under review)

# Knowledge Resources

Sources of exploitable structured human knowledge:

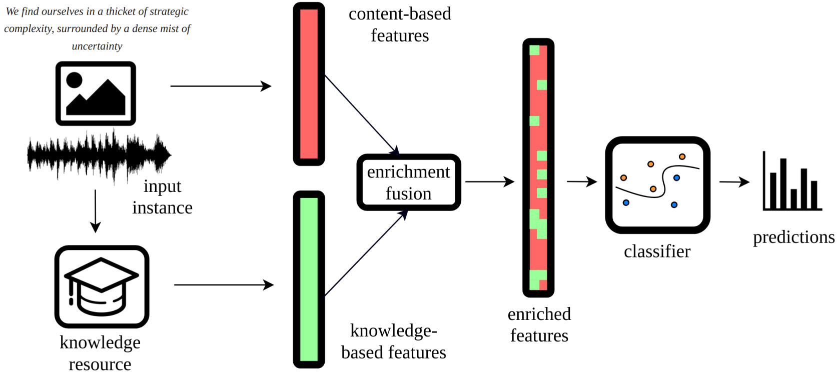
- Semantic Graphs (Wordnet, Framenet, ConceptNet)
- Property-value stores (DBpedia, Wikidata)
- Lexicons (E-ANEW, GeneralInquirer)
- Hierarchical labelsets / ontologies (Imagenet, Audioset)

How do we use knowledge resource?

- Retrieve relevant knowledge per instance
- Integrate knowledge based on enrichment method

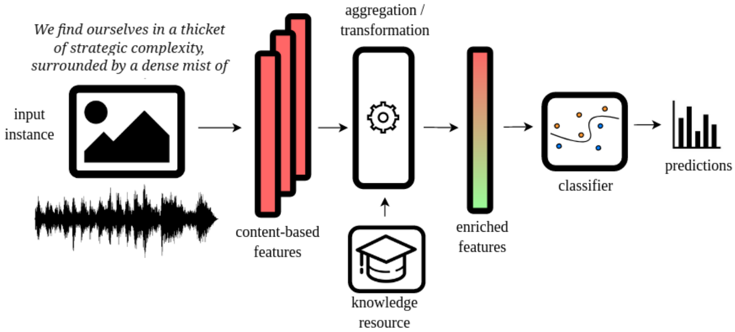
# Input enrichment / modification

- Augment feature set from content-based methods
- Inject knowledge-based features in the representation
- Result: discrete, joint content + knowledge-based feature space



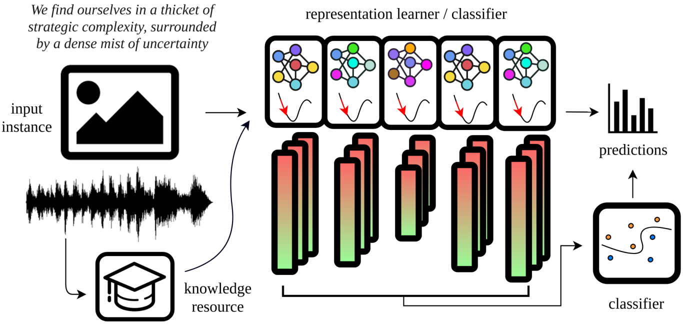
# Knowledge-based refinement

- Transform / combine / aggregate content-based features
- Refinement guided / oriented / informed via knowledge
- Distributed enriched features



# Knowledge-aware deep systems

- Hierarchical, deep task / representation end-to-end learners
- Ingest content-based and knowledge-based information
- Enrichment process learned jointly with the representation



# Findings

Multiple, diverse enrichment avenues in the literature:

- Correspondence to content-based paradigms
- Similar strengths and weaknesses apply
  - Low-level/template-matching → input modification/enrichment
  - Aggregation-based → knowledge-based refinement
  - Deep rep. learners → end-to-end knowledge-aware systems

→ **Can we select and utilize the best elements per paradigm?**

# Proposed Approach

Proposed approach:

- Explore promising combination not explored in the literature:

1. Enrichment of **deep content-based features**
2. Use the **input modification enrichment**

Combine strengths:

- Rich, expressive content-based features
- Intuitive, explainable enriched representation

# Overview

Proposed enrichment approach [1]:

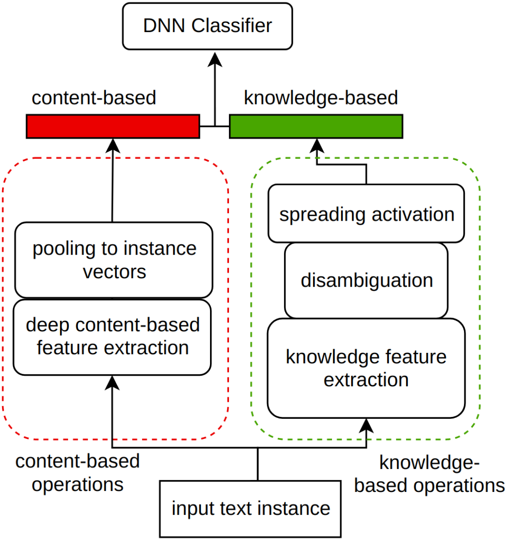
- Word embedding features
- Disambiguated word senses from semantic graph
- Input enrichment / modification
- Exploit knowledge resource structure via spreading activation
- Deep Neural Network classifier

[1] Pittaras et al., Text classification with semantically enriched word embeddings, NLE Special Issue: Informing Neural Architectures for NLP with Linguistic and background Knowledge

# Contributions

- Implementation of knowledge-enriched classification system
- Large-scale, cross-domain, comparative empirical evaluation
- Verification of performance benefits of proposed enrichment
- Statistically significant results
- State of the art results
- Identification of future directions for improvement

# Overview


DNN Classifier
content-based
knowledge-based
pooling to instance vectors
deep content-based feature extraction
spreading activation
disambiguation
knowledge feature extraction
content-based operations
knowledge-based operations
input text instance

# Content-based features

Content-based component:

- Neural word embeddings
- CBOW model (Word2Vec, (Mikolov, 2013a))
- 50-epoch training, 10-word window
- Average word vectors to document representations

# Knowledge Resource

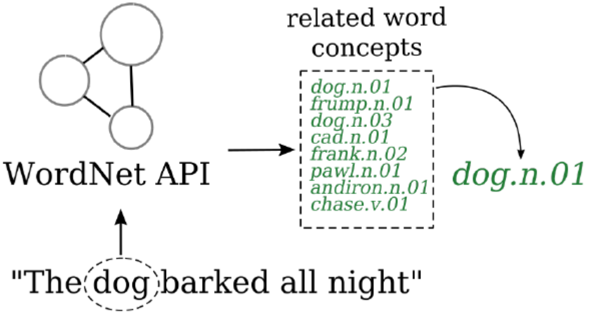
Knowledge Resource:

- WordNet v3 (Miller, 1995) semantic graph
- Built from sense-annotated SemCor corpus (Landes, 1998)
- Nodes: set of synonymous word senses (Synsets)
- Edges: hyponymy, meronymy, hypernymy, etc. relations
- POS information, lexical literals per sense
- Mine sense information from words in the text

# Knowledge Extraction - Basic

WordNet information retrieval

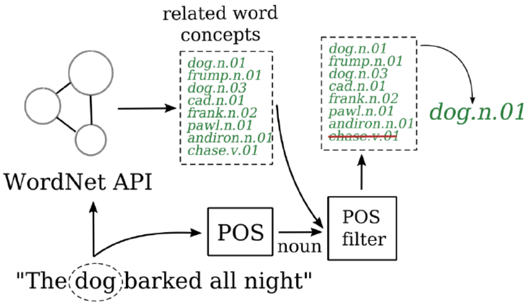
- Disambiguation required for multisense words
- Senses extracted sorted by frequency in WordNet corpus[1]
- "Basic" disambiguation: retrieve the most common sense




[1] wrt. NLTK Wordnet API

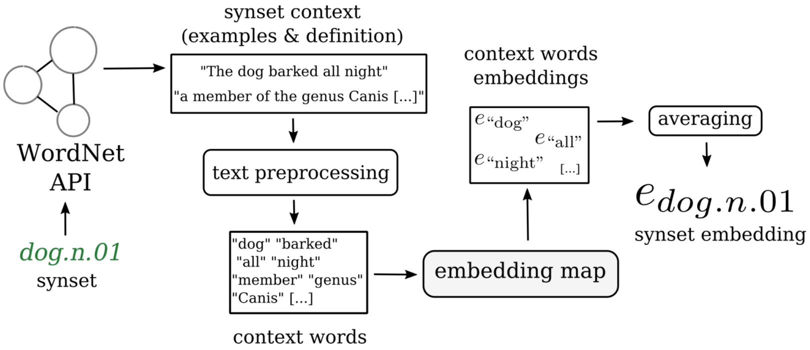
# Knowledge Extraction - POS

- E.g. Senses for *slack*
  - verb: to avoid responsibility / work
  - noun: deterioration in performance
  - adj: loose, not taught
  - . . .
- Extract senses for input word, filter to match input POS
- Proceed with "Basic" disambiguation



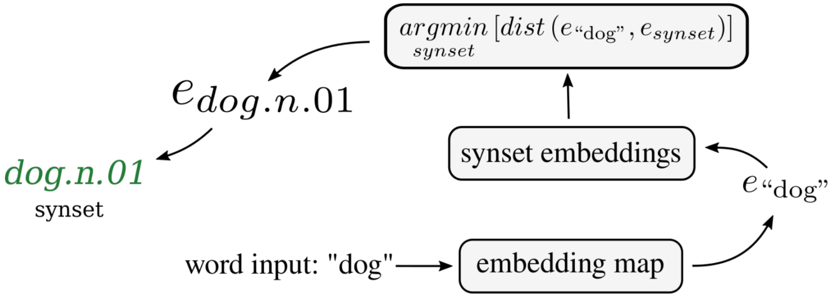
# Knowledge Extraction - Semantic embeddings

- Build synset vectors from their context (definition & examples)
- Aggregate context to vectors (as in the content-based case)
- Use resulting vector as sense representative
- Lies in the space of content-based embeddings



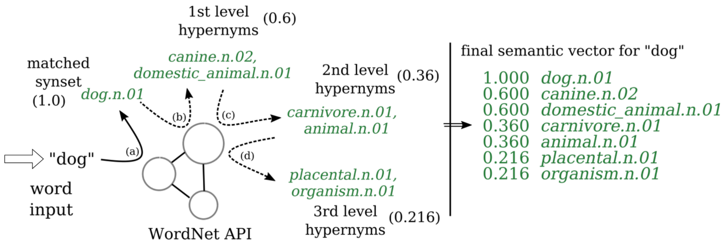
# Knowledge Extraction - Semantic embeddings

- For an input word, use its content-based embedding
- Select synset having an embedding with the closest distance



# Knowledge Extraction - Spreading Activation

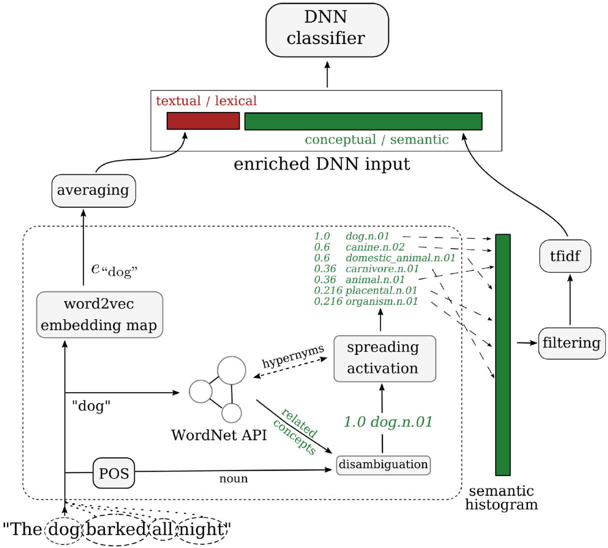
- Exploit Wordnet hypernymy structure
- For a extracted synset, also recursively use its parents
- Decay activation (weight) of match with each propagation



# Proposed Configurations

- BoW / TF-IDF semantic vectors
- Keep all / top $K$ / with min. freq. $K$ senses
- Basic / POS / semantic-embedding disambiguation
- With / without spreading activation
- Concatenate with / replace content-based features

# Overview


DNN
classifier
textual / lexical
conceptual / semantic
enriched DNN input
averaging
$e_{\text{“dog”}}$
word2vec
embedding map
1.0 dog.n.01
0.6 canine.n.02
0.6 domestic_animal.n.01
0.36 carnivore.n.01
0.36 animal.n.01
0.216 placental.n.01
0.216 organism.n.01
tfidf
filtering
spreading
activation
hypernyms
"dog"
WordNet API
related
concepts
1.0 dog.n.01
disambiguation
POS
noun
semantic
histogram
"The dog barked all night"

# Datasets

- 20Newsgroups: USENET forum posts, 20 labels
- Reuters-21578: Reuters financial articles, 90 labels
- Different domains, text / labelset sizes
- Balanced vs. imbalanced
- Different deegrees of useful POS / Wordnet information

| | 20-Newsgroups | | Reuters | |
|---|---|---|---|---|
| attribute | train | test | train | test |
| samples | 11,314 | 7,532 | 9,584 | 3,744 |
| class samples | 377 - 600 | 251 - 399 | 1 - 2,877 | 1 - 1,087 |
| words | 191.164 (587.7) | 172.196 (471.37) | 92.532 (92.03) | 92.899 (105.25) |
| POS | 0.716 (0.07) | 0.713 (0.06) | 0.672 (0.10) | 0.669 (0.10) |
| WordNet | 0.572 (0.09) | 0.566 (0.09) | 1.479 (0.37) | 1.381 (0.38) |

# Experimental Setup

- Feed-forward DNN, tuned to 2 layers and 512 neurons
- 50-epoch training with early stopping and LR decay
- 5-fold cross-validation, significance testing
- Mi/ma/per-class F1-score
- Implementation with python3, keras, tensorflow, Wordnet v3

# Experimental Results

## State of the art performance

| config | Reuters | | 20-Newsgroups | |
|---|---|---|---|---|
| system | accuracy | ma-f1 | accuracy | ma-f1 |
| majority baseline | 0.290 | 0.005 | 0.005 | 0.053 |
| embedding-only | 0.725 | 0.295 | 0.724 | 0.716 |
| our approach | **0.749** | **0.378** | **0.784** | **0.790** |
| other trained embeddings | | | | |
| FastText [Joulin et al., 2017] | <u>0.732</u> | <u>0.319</u> | <u>0.751</u> | <u>0.743</u> |
| FastText + retrofitting | 0.717 | 0.260 | <u>0.748</u> | <u>0.740</u> |
| word2vec + retrofitting | 0.709 | 0.248 | 0.717 | 0.710 |
| pre-trained embeddings | | | | |
| glove [Pennington et al., 2014] | 0.702 | 0.275 | 0.620 | 0.610 |
| glove + retrofitting | 0.684 | 0.235 | 0.587 | 0.575 |
| FastText | <u>0.733</u> | <u>0.310</u> | <u>0.734</u> | <u>0.727</u> |
| FastText + retrofitting | 0.705 | 0.239 | 0.706 | 0.695 |
| word2vec (300-dim) | <u>0.737</u> | <u>0.311</u> | 0.721 | 0.712 |
| word2vec (300-dim) + retrofitting | 0.689 | 0.239 | 0.476 | 0.465 |
| single-context [Huang et al., 2012] | 0.661 | 0.227 | 0.541 | 0.531 |
| single-context + retrofitting | 0.629 | 0.175 | 0.464 | 0.454 |
| pre-trained sense embeddings | | | | |
| multi-context [Huang et al., 2012] | 0.570 | 0.121 | 0.430 | 0.412 |
| SensEmbed [Iacobacci et al., 2015] | <u>0.728</u> | <u>0.308</u> | 0.722 | 0.714 |
| Supersenses [Flekova and Gurevych, 2016] | <u>0.729</u> | <u>0.313</u> | <u>0.733</u> | <u>0.725</u> |

# Error Analysis

Prevalent error cases (e.g., 20Newsgroups)

- Semantically similar labels (religion, atheism, christianity)
- Ambiguous / equivocal instances ("Abortion government funding": religion / politics)
- Critical named-entities ("Jack Morris" / baseball, VAX / computer )
- Context misses ("The devil reincarnate": autos / religion)

$\rightarrow$ Important finding: **Explainable / edge-case / intuitive errors**

# Additional datasets / domains

| | bbc | | ohsumed | |
|---|---|---|---|---|
| system | accuracy | ma-f1 | accuracy | ma-f1 |
| majority | 0.230 | 0.075 | 0.172 | 0.013 |
| embedding-only | 0.970 | 0.970 | 0.384 | 0.300 |
| ours | **0.976** | **0.976** | **0.435** | **0.373** |
| other pre-trained embeddings | | | | |
| word2vec | 0.973 | 0.973 | 0.307 | 0.244 |
| word2vec + retrofitting | 0.880 | 0.878 | 0.313 | 0.250 |
| SensEmbed | 0.969 | 0.969 | 0.328 | 0.215 |
| Supersenses | 0.852 | 0.851 | 0.229 | 0.148 |

# Findings

Overall:

- Enrichment: high, statistically significant performance boost
- State of the art results on multiple datasets and domains
- Explainable, intuitive errors

Proposed configurations:

- Context embeddings show poor performance (thresholds)
- Concatenating works best: content is valuable
- TF-IDF outperformed by count-based semantic vectors
- Spreading activation contribution varies across datasets
- Sem. vectors reduced by 61% retain 99.36% of performance
  - Suggestion: dimensionality reduction for semantic features

# Summary of Contributions

- Investigate unexplored enrichment combination

1. Enrichment of **deep content-based features** ✓
   → **Word2Vec word embeddings**
2. Use the **input modification** enrichment ✓
   → **Concatenation / replacement with WordNet sense-based information**
   → **Utilize the architecture of the knowledge resource**

# Summary of Contributions

- Implementation of knowledge-enriched classification system
- Large-scale, cross-domain, comparative empirical evaluation
- Proposed enrichment gives statistically significant improvements
- State of the art results
- Identification of directions for future work

# Motivation

Motivation of the proposed method [1]:

- Based on stated goals and previous findings
- Evaluate proposed enrichment in additional task
- Examine the enrichment of other embedding methods
- Investigate dimensionality reduction of enriched features

[1] Pittaras et al., A study of semantic augmentation of word embeddings for extractive summarization, Multiling 2019

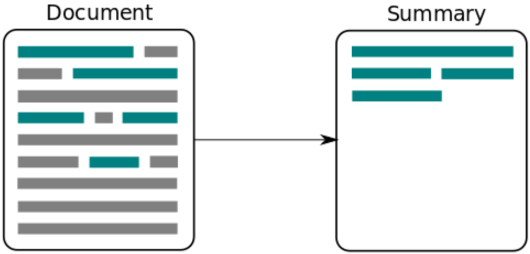
# Contributions

Investigation and evaluation of:

- Proposed enrichment in the summarization task
- Enrichment of different deep content-based features
- Dimensionality reduction of enriched information:
  - via different / diverse reduction methods
  - arriving at different reduced dimensionalities
  - applying reduction at different stages in enrichment

# Extractive Summarization

- Extractive: retain important sentences from source text
- Arrive to a cohesive, informative summary
- Enrichment focus: classification for sentence selection



# Enriched Representation

Content-based information:

- CBOW model (Word2Vec), as used previously
- Pretrained subword embeddings (FastText (Joulin, 2016))
- TF-IDF baseline

Semantic enrichment:

- Wordnet semantic features (Miller, 1995)
- "Basic" disambiguation strategy
- Concatenation to the content-based vector

# Dimensionality Reduction

Diverse selection of established methods:

- Principal Component Analysis (PCA) (Jollife, 2011a)
  - Transformation with respect to feature variance
- Latent Semantic Analysis (LSA) (Deerwester, 1990)
  - Feature decomposition to latent topics
- K-Means clustering (Lloyd, 1982)
  - Distance-based grouping

# Dataset

- Multiling 2015 Single-Document Summarization Dataset (Giannakopoulos, 2015)
- English Wikipedia articles & summaries
- Sentence-level annotation (1: include in summary, 0: don't) based on ranked ngram overlaps between source / summary
- Severely imbalanced, arrived at oversampling to 2 : 1 for training

| feature | train | test |
|---|---|---|
| document sentences | 233 | 184.9 |
| document summary sentences | 77.9 | 13.5 |
| document words | 25.5 | 22.8 |
| samples | 6990 | 5546 |

# Proposed Approaches and Experimental Setup

Content-based information

- CBOW-Word2Vec (50-dim) / FastText (300-dim) / TF-IDF

Dimensionality reduction:

- PCA, LSA or KMeans
- Evaluate reductions to 50, 100, 250, 500 dimensions
- Apply only on knowledge features or the entire enriched vector

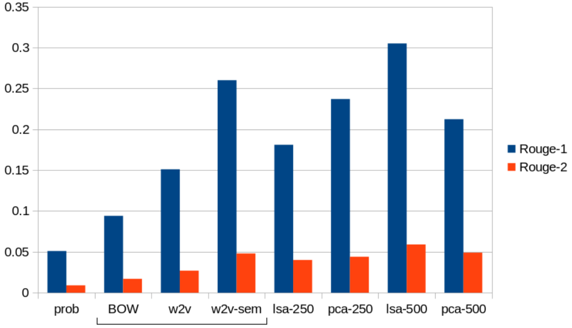
Learning model

- Feed-forward $5 \times 512$ DNN, 5-fold CV
- Rely on Rouge 1 & 2 for evaluation

# Experimental Results

Content-based and enriched features:

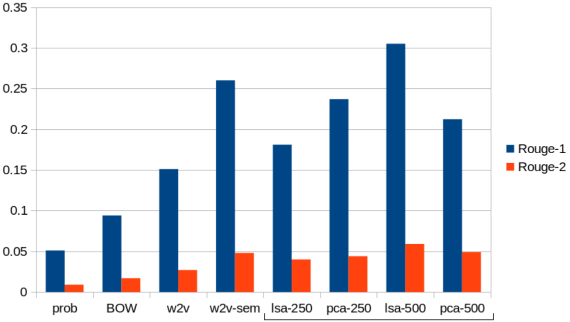
- Word2Vec (shown) and FastText perform similarly
- BOW < embeddings < enriched embeddings
- Enrichment: improves summarization performance
- Encourages selection of informative sentences



# Experimental Results

Effect of dimensionality reduction methods:

- concatenate then reduce (shown) > reduce then concatenate
- Reduction can improve summarization performance
- PCA features most robust to severe reductions
- LSA > PCA >> KMeans



# Findings

General:

- BOW < embeddings < enriched embeddings
- Word2Vec ≈ FastText perform similarly

Effect of semantic enrichment:

- Encourages selection of informative sentences
- Improves summarization performance

Effect of dimensionality reduction methods:

- Reduction can improve summarization performance
- Performance improves with less reduction, PCA most robust
- LSA > PCA >> KMeans
- concatenate, reduce enriched > reduce knowledge, concatenate

# Contributions

- Proposed enrichment in the summarization task ✓
  → **Verified improvement over content-based baselines**
- Enrichment of different deep content-based features ✓
  → **Examined FastText alternative**
- Dimensionality reduction of enriched information ✓
  - via different / diverse reduction methods
  - arriving at different reduced dimensionalities
  - applying reduction at different stages in enrichment

  → **Investigated different of LSA, PCA, KMeans in different configurations**

# Summary of Representation Enrichment

- Multiple avenues for knowledge-based enrichment
- Proposal: input modification/enrichment of deep features
  - Rich learned semantics with explainable high-level knowledge
  - Classification: state of the art performance
  - Summarization: improves content-based approaches, amenable to dim. reduction

# Industrial Application

# Acknowledgements

Athens Technology Center (ATC[1]): Stavros Niarchos Industrial Scholarship partner

[1] https://www.atc.gr/

# Requirements

Utilize findings on representation enrichment for [16]:

- Use case: Hate Speech Detection / Multiclass Classification
- Real-world deployment

Desired features:

- Easy deployment, fine-tuning and monitoring
- Easy extension / maintenance

# Data

- English Social media short, noisy texts
- Combination of existing HSD datasets + data crawling
- Domain-specific preprocessing
- HS type classes: racism, sexism, misogyny, religious, none

| label | train | | test | |
|---|---|---|---|---|
| | # instances | mean # words | # instances | mean # words |
| racism | 2448 | 14.22 | 15 | 14.0 |
| sexism | 4213 | 15.57 | 15 | 17.53 |
| orientation | 677 | 12.78 | 15 | 12.47 |
| religion | 581 | 19.18 | 15 | 20.0 |
| none | 7761 | 13.99 | 15 | 16.53 |
| overall | 15680 | 14.59 | 75 | 16.11 |

# Representations approaches

Content-based:

- Word2Vec & FastText embeddings
- Bag of Words

Enrichment:

- Bag of Semantic Units
- WordNet hypernym information
- Compiled list of hateful keywords / phrases

# Learning and Tuning

Learning models:

- Feedforward DNN
- Logistic Regression
- Under/over-sampling functionality

Tuning:

- Scalable hyperparameter tuning with ray tune (Liaw, 2018)
- Large-scale grid search

# Monitoring and Deployment

Model monitoring / comparison:

- MLFlow MLops tool (Zaharia 2018)

Deployment:

- MLFlow
- Flask, Swagger (Grinberg 2018, De 2017)

Implementation:

- python 3.8, based on the numpy / sklearn stack
- Domain-specific packages for crawling, preprocessing, etc.

# Contributions

- Extensible, optimized Hate Speech Detection system
- Utilization of state of the art in representation enrichment
- Utilizing modern approaches in MLOps

# Conclusion

# Contributions

Content-based representations:

- Comparative literature review
    - Organization with respect to representation sophistication
    - → **Verified motivation for pursuit of rich, expressive features**
- Proposal of novel approaches and applications
    - Vector-based and graph-based representations
    - Classification, summarization, clustering tasks
    - Text, image and audio data modalities
    - → **Very difficult / complex to discover one universally optimal approach**

# Contributions (cont.)

Representation enrichment with external knowledge:

- Comparative literature review on enrichment
  - Organization with respect to enrichment type
  - Detailed presentation of knowledge resources
  - → **Identified under-investigated approaches in the literature**
- Proposal of novel enrichment strategies
  - Input enrichment of deep features with WordNet semantics
  - → **Large-scale investigation on text classification, SotA results**
  - Extension with dim. reduction and additional deep features
  - → **Investigation on text summarization, verifying improvements**
- Utilization of conducted research in an industrial setting

# Misc. Contributions

Academic activities during the project:

- Support work for DiT-UoA (exams / courses)
- Reviewing for journals, conferences and workshops (e.g. Machine Learning, CSL, ICTAI)
- Co-organization of conferences and workshops (e.g. SETN2020, FNP/FNS 2020, 2021)
- Contribution / creation of relevant open-source software (e.g. JINSECT)
- BSc. / MSc. student theses co-supervision

# Findings

Key take-aways:

- High-level representation semantics crucial for usefulness in downstream tasks
- Representation has significant impact on learning for different tasks / modalities
- Improvable with high-quality human knowledge
- Proposal effectively exploits deep learning features and conceptual information
- Improves the state of the art by applying representation enrichment

# Future Work

Proposed approach:

- Additional knowledge resources
- Combination of multiple knowledge resources
- Dimensionality reduction with representation learning (e.g. autoencoders, FeedForward networks)

Representation enrichment:

- Development of easy-to-use knowledge resources for modalities other than text
- Combination of multiple enrichment strategies (e.g. input modification followed by refinement)

# List of publications

**Conferences**

1 N. Pittaras, V. Karkaletsis, "A study of semantic augmentation of word embeddings for extractive summarization", Multliling Workshop, RANLP2019, Varna, Bulgaria.

2 N. Gialitsis, N. Pittaras, P. Stamatopoulos, "A topic-based sentence representation for extractive text summarization", Multliling Workshop, RANLP2019, Varna, Bulgaria.

3 C. Themeli, G. Giannakopoulos, N. Pittaras "A study of text representations for Hate Speech Detection", CICLING 2019, La Rochelle, France.

4 N. Pittaras, G. Papadakis, G. Stamoulis, G. Argyriou, E. K. Taniskidou, E. Thanos, G. Giannakopoulos, L.Tsekouras, E. Koubarakis, "GeoSensor: Semantifying Change and Event Detection over Big Data", SAC 2019, Limassol, Cyprus.

5 A. Kosmopoulos, A. Liapis, G. Giannakopoulos, N. Pittaras. (2020). Summarizing Game Reviews: First Contact. SETN Workshops.

6 M. El-Haj, M. Litvak, N. Pittaras, G. Giannakopoulos. "The Financial Narrative Summarisation Shared Task (FNS 2020)." Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation. 2020.

7 G. Papadakis, L. Tsekouras, M. Thanos, N. Pittaras, G. Simonini, D. Skoutas, P. Isaris, G. Giannakopoulos, T. Palpanas, M. Koubarakis. "JedAI3: beyond batch, blocking-based Entity Resolution." EDBT. 2020.

8 G. George, N. Pittaras. "The Summary Evaluation Task in the MultiLing-RANLP 2019 Workshop." Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources. 2019.

9 N. Pittaras, G. Giannakopoulos, L. Tsekouras, I. Varlamis. "Document clustering as a record linkage problem." Proceedings of the ACM Symposium on Document Engineering 2018. 2018.

# List of publications

**Journals**

10 N. Pittaras, G. Giannakopoulos, G. Papadakis, V. Karkaletsis,“Text classification with semantically enriched word embeddings”, Natural Language Engineering Special Issue: Informing Neural Architectures for NLP with Linguistic and background Knowledge

11 N. Pittaras, G. Giannakopoulos, P. Stamatopoulos, V. Karkaletsis , “Content-based and knowledge-enriched representations for classification across modalities: a survey”, ACM Computing Surveys (submitted, under review)

12 N. Pittaras, T. Giannakopoulos, S. Perantonis, “A study of deep audio-visual fusion methods for video classification” (journal paper, to be submitted)

**Book Chapters**

13 G. Giannakopoulos, G. Kiomourtzis, N. Pittaras, V. Karkaletsis. Scaling and Semantically-Enriching Language-Agnostic Summarization. In A. Fiori (Ed.), Trends and Applications of Text Summarization Techniques (pp. 244-292). IGI Global, 2020.

14 N. Pittaras, S. Montanelli, G. Giannakopoulos, A. Ferrara, V. Karkaletsis. "Crowdsourcing in single-document summary evaluation: the argo way." Multilingual Text Analysis: Challenges, Models, and Approaches. 2019. 245-280.

**Tech. reports**

15 N. Pittaras, N. Kostagiolas, C. Nikolaou, G. Giannakopoulos. "Exploring different sequence representations and classification methods for the prediction of nucleosome positioning." bioRxiv (2018): 482612.

16 N. Pittaras, G. Giannakopoulos, V. Karkaletsis, “Enriched Representations for Hate Speech Detection” (technical report, to be finalized)

# Thank you

# Appendix

## Hate Speech Detection

Modality / Task:

- Text, Classification

Approaches:

- BoW, N-gram Graphs, GloVe Embeddings, syntax, spelling
- Different classifiers (KNN, LR, NB, MLP, RF)

Findings:

- Representation statistically more significant than classifier
- GloVe word embeddings achieve best performance
- N-gram graph representations produce rich features

# Extractive Summarization of Web Documents

Modality / Task:

- Text, Automatic Summarization

Approaches:

- Modeled as binary sentence classification
- Topic-based vs shallow features (LDA, TF-IDF)
- Different classifiers (DT, KNN, GB, NB, LiDA, QDA, LR, SVM)

Findings:

- LDA topic-based method produces robust features
- Improvement over the TF-IDF-based classification

# Automatic Summarization of Video Game Reviews

Modality / Task:

- Text, Automatic Summarization

Approaches:

- Multiple aspect identification & labelling pipelines
- K-Means clustering, keyword matching, sentiment analysis
- Evaluated with feedback from human surveys
- TF-IDF and BERT representations with LR classifiers
- NewSum for extractive sentence selection

Findings:

- No clear winner between evaluted representations
- Verified aspect extraction as a crucial step
- Identified unique challenges for the domain of game reviews

# Clustering, Summarization and Classification of Web Documents and Social Media in the Security Domain

Modality / Task:

- Text, Classification, Clustering, Automatic Summarization

Approaches:

- N-Gram Graphs for text / social media representations
- Similarity-based clustering, summarization, classification
- Integration with multi-purpose platform operating on diverse big data sources

Findings:

- Graph-based approaches can provide rich representations
- Identified performance bottlenecks on similarity extraction

## Scaling and Enrichment of Automatic Summarization

Modality / Task:

- Text, Automatic Summarization

Approaches:

- Expand graph-based text representations (e.g. with NER)
- Similarity extraction by distributed execution (SPARK)

Findings:

- Considerable acceleration via SPARK-based operations
- Identified optimal speedup / hardware trade-offs

# Classifying Videos with Multimodal DNNs

Modality / Task:

- Video (image and audio), Classification

Approaches:

- FeedForward / LSTM networks for handling temporal relations
- Audio / visual modalities, multimodal configurations
- Multiple, diverse video datasets and domains

Findings:

- LSTMs outperformed by FF nets on audio and vice versa
- Considerable impact of the modality and domain
- Weighted linear combination of single-modality works best
- Deep representations outperform engineered features

# Similarities to Content-based Paradigms

Input enrichment / modification

- Resemblance to low-level / template-matching methods
- Knowledge handled as distinct data coordinates
- Explainable, discrete features

Knowledge-based refinement

- Resemblance to aggregation methods
- Aggregation mechanism defined, parameterized by knowledge
- Mostly explainable, distributed features

Knowledge-aware end-to-end systems

- Resemblance to deep representation learning
- Jointly learn to consider knowledge along with content
- Non-explainable, distributed features