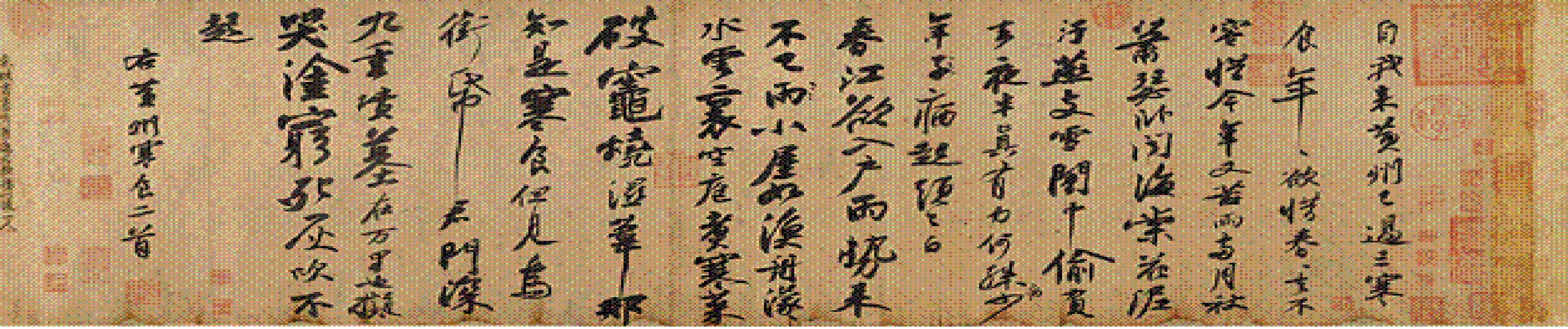


信息检索
Information Retrieval

教师：孙茂松

Tel:62781286

Email:sms@tsinghua.edu.cn

TA：胡锦毅

Email:hu-jy21@mails.tsinghua.edu.cn

郑重声明

- 此课件仅供选修清华大学计算机系本科生课《信息检索》（课号：40240372）的学生个人学习使用，所以只允许学生将其下载、存贮在自己的电脑中。未经孙茂松本人同意，任何人不得以任何方式扩散之（包括放到9#服务器上）。否则，由此可能引起的一切涉及知识产权的法律责任，概由该人负责。
- 此课件仅限孙茂松本人讲课使用。除孙茂松本人外，凡授课过程中，PTT文件显示此《郑重声明》之情形，即为侵权使用。

第九章 文本自动分类

Yahoo!

Yahoo! - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
地址(D) http://www.yahoo.com/
转到 链接
Web Site Directory - Sites organized by subject
Suggest your site
Business & Economy
B2B, Finance, Shopping, Jobs...
Regional
Countries, Regions, US States...
Computers & Internet
Internet, WWW, Software, Games...
Society & Culture
People, Environment, Religion...
News & Media
Newspapers, TV, Radio...
Education
College and University, K-12...
Entertainment
Movies, Humor, Music...
Arts & Humanities
Photography, History, Literature...
Recreation & Sports
Sports, Travel, Autos, Outdoors...
Science
Animals, Astronomy, Engineering...
Health
Diseases, Drugs, Fitness...
Social Science
Languages, Archaeology, Psychology...
Government
Elections, Military, Law, Taxes...
Reference
Phone Numbers, Dictionaries, Quotations...
Buzz Index - Yahoo! Picks - New Additions - Full Coverage
powered by hp
Up to $100 off with rebate. Free shipping and 18 months financing at SonyStyle
Free month of Vonage phone service - Calling plans start at only $14.99/month.
Hottest holiday toy - Get Leapster from LeapFrog online at Wal-Mart
Digital camera only $59.99 ($170 value) - With any $150 online purchase at Macy's
Find Yu-Gi-Oh cards on Yahoo! Auctions
Shopping - Computers - Electronics - Travel
Entertainment
Survivor Insider - Exclusive Video
Sandra questions Christa's loyalty in an unaired scene from last week's episode. Watch it now
LAUNCH - Year In Music 2003
Watch 2003's hottest performances, plus check out the best videos and vote for the song of the year.
Entertainment - Games - Movies - Music - TV

Yahoo!

Yahoo! Directory Computers and Internet - Microsoft Internet Explorer

文件 (F)　编辑 (E)　查看 (V)　收藏 (A)　工具 (T)　帮助 (H)

地址 (D) http://dir.yahoo.com/Computers_and_Internet/

CATEGORIES

Commercial Categories

- Business to Business@
- Consumer Products and Services@

Additional Categories

- Communications and Networking (1150) NEW!
- Computer Generated Art@
- Computer Generated Music@
- Computer Science@
- Computers and Technology Weblogs@
- Contests (20)
- Conventions and Trade Shows@
- Countries, Cultures, and Groups (38)
- Cyberculture@
- Data Formats (491) NEW!
- Desktop Customization@
- Desktop Publishing (43)
- Dictionaries (25)
- Employment@
- Ethics (14)
- Forensics (11)
- Games@
- Graphics (323)
- Hardware (4069)
- History (111)
- Humor@
- Industry Information@
- Information Technology@
- Issues (9)
- Macintosh (486)
- Magazines@
- Mobile Computing (59)
- Multimedia (588)
- News and Media (196)
- Operating Systems@
- Organizations (87)
- People (194)
- Personal Digital Assistants (PDAs) (626) NEW!
- Product Information and Reviews (246)
- Programming and Development (1392)
- Science and Technology Policy@
- Security and Encryption (652)
- Software (3841) NEW!
- Standards (36)
- Supercomputing and Parallel Computing@
- Technical Guides and Support (68)
- Telecommunications@
- Training@
- User Groups@
- Virtual Reality (105)
- Web Directories (16)

Definition

… the activity of labeling natural language texts with thematic categories from a pre-defined set

[Sebastiani, ACM Computing Surveys, 1-47,2002]

Also called “Text Classification,” “Topic Spotting”

NLP in TC: Sample Image with Caption

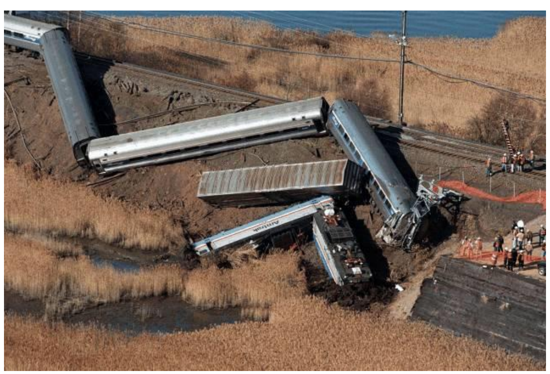
Philippine rescuers carry a fire victim March 19 who perished in a blaze at a Manila disco.

Categorizing images based on captions

Categories

Politics

Struggle



Disaster

Crime

Other

Categories (cont)

Politics

Struggle

Disaster

Crime

Other

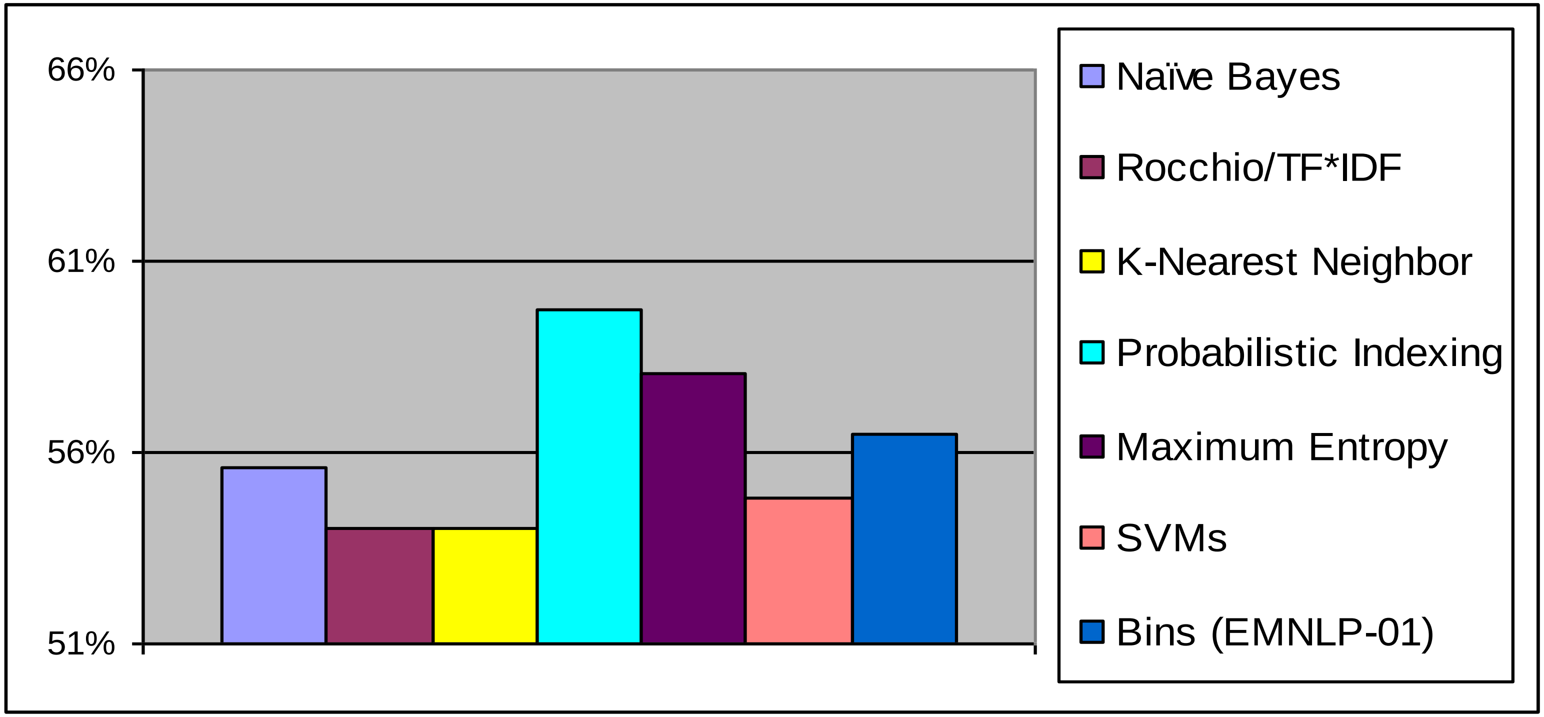
Affected People

Workers Responding

Wreckage

Other

Performance of Standard Systems Not Very Satisfying



Words are Ambiguous:
Workers Responding vs. Affected People

Philippine rescuers carry a fire victim March 19 who perished in a blaze at a Manila disco.

Workers Responding

Affected People

Hypothetical alternative caption: *A fire victim who perished in a blaze at a Manila disco is carried by Philippine rescuers March 19.*

Observations About the Task

Philippine rescuers carry a fire victim March 19 who perished in a blaze at a Manila disco.

- Need to distinguish foreground from background, determine focus of image
- Not all words are important; some are misleading
  - Problematic for bag of words approaches
- Hypothesis: subject and verb are useful clues
  - Need linguistic analysis to determine predicate argument relationships

Hypothesis: Subject and Verb are Useful Clues

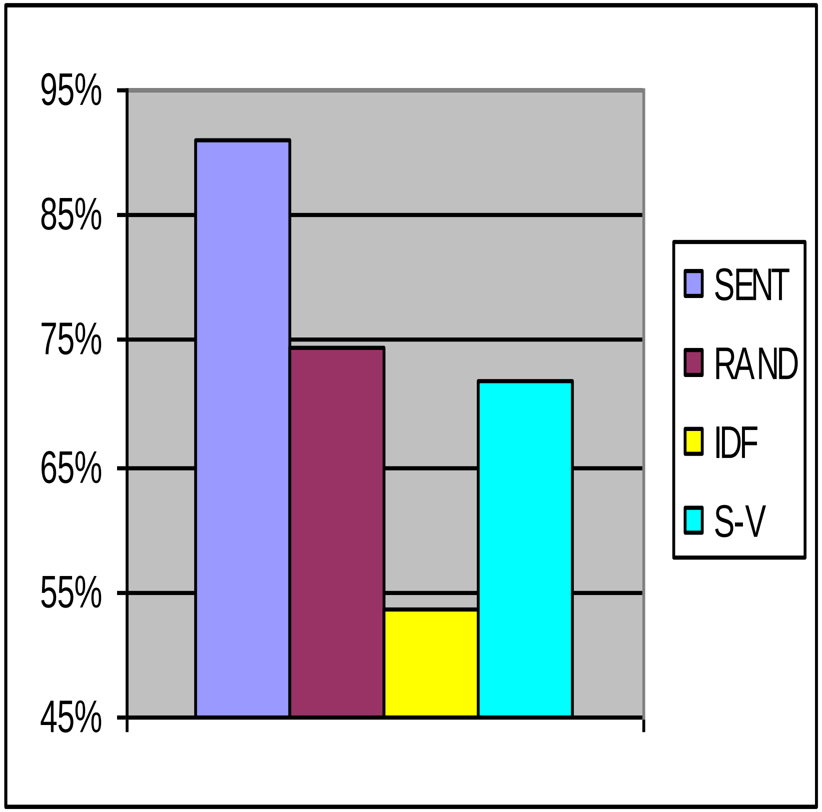
Subject	Verb	Category	Guessable?
couple	mourn	Affected People	**Yes**
NAME	gather	Affected People	No
inspectors	search	Workers Responding	**Yes**
NAME	observes	Workers Responding	No

Experiments with Humans Subjects: 4 Conditions

Test Hypothesis: Subject and Verb are Useful Clues

SENT: First sentence of caption	Philippine rescuers carry a fire victim March 19 who perished in a blaze at a Manila disco.
RAND: All words from first sentence in random order	At perished disco who Manila a a in 19 carry Philippine blaze victim a rescuers March fire
IDF: Top two TF*IDF words	disco rescuers
S-V: Subject and verb	subject = “rescuers”, verb = “carry”

Experiments with Humans Subjects: Results
Hypothesis: Subject and Verb are Useful Clues



- More words are better than fewer words
  - SENT, RAND > S-V, IDF
- Syntax is important
  - SENT > RAND; S-V > IDF

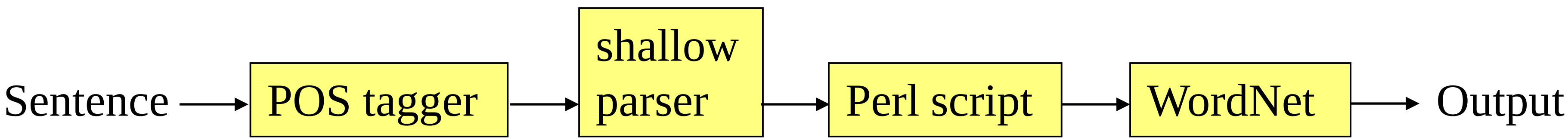
RAND is Very Slow!

Condition	Average Time (in seconds)
RAND	68
SENT	34
IDF	22
S-V	20

Perhaps human subjects unscrambled words, regaining syntactic information

Operational NLP-based System

- Extract subjects and verbs from all documents in training set



For each test document:

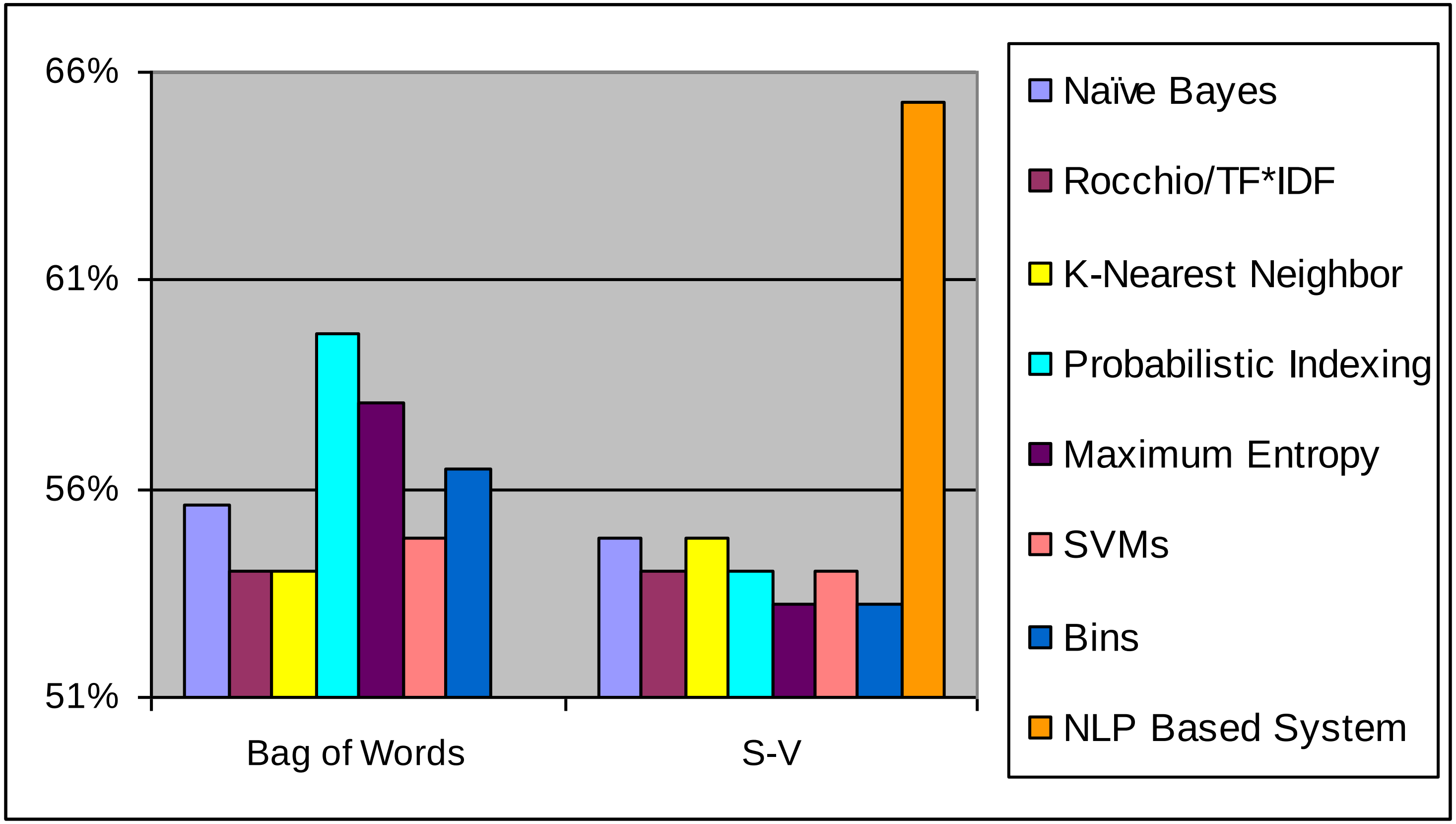
Extract subject and verb

Compare to those from training set using some method of word-to-word similarity

Based on similarities, generate a score for every category

NLP-based System Outperforms Others

The Right Two Words Beat All the Words,
NLP Found Helpful for at least one Text Categorization Task!



NLP is Important for the Task!

- Not all words are important; some are misleading
- Need to distinguish foreground from background, determine focus of image
- Subject and verb: clues for focus

Verified in two ways:

- Experiments with human subjects
- Operational NLP-based system outperforms others

Philippine **rescuers carry** a fire victim March 19 who perished in a blaze at a Manila disco.

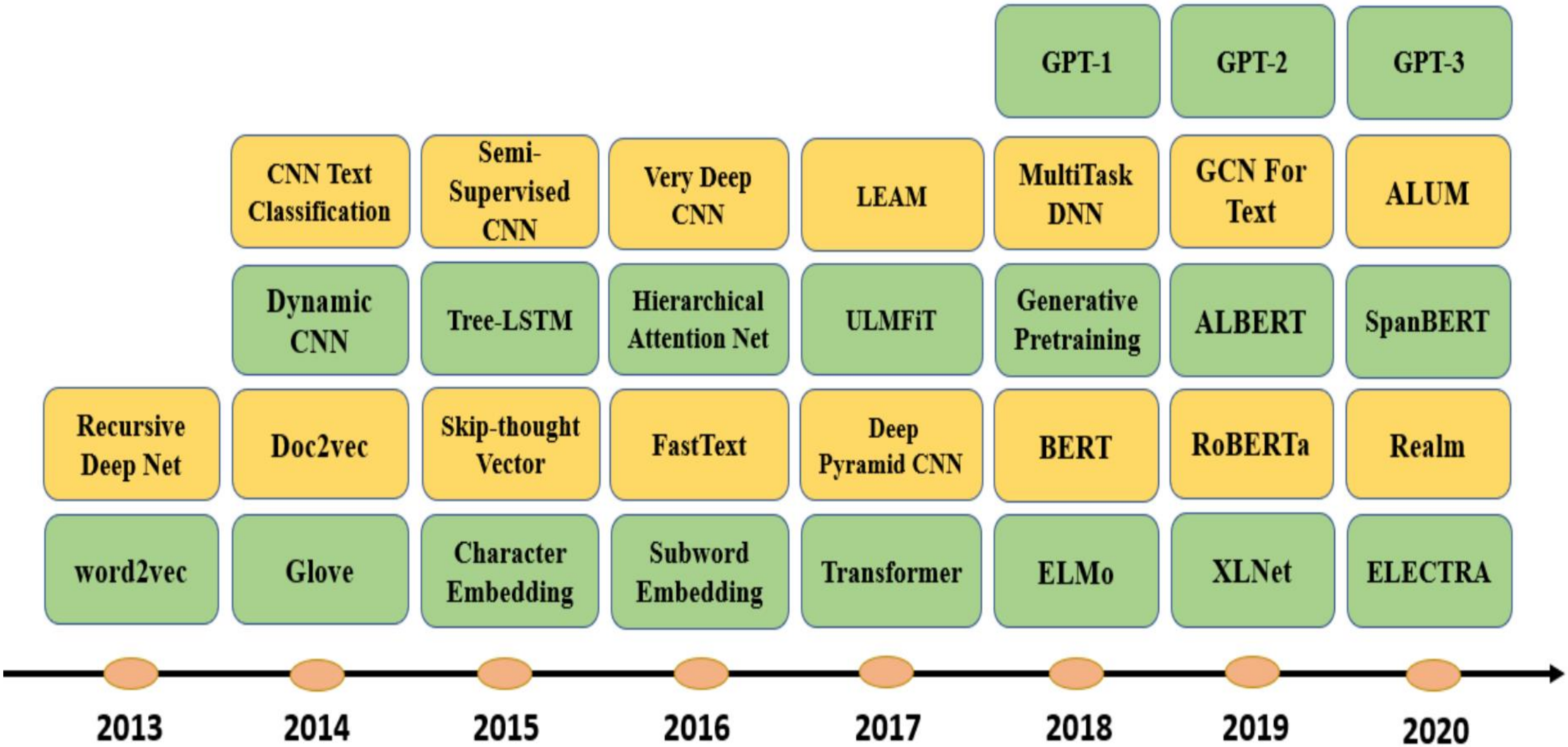

GPT-1
GPT-2
GPT-3
CNN Text Classification
Semi-Supervised CNN
Very Deep CNN
LEAM
MultiTask DNN
GCN For Text
ALUM
Dynamic CNN
Tree-LSTM
Hierarchical Attention Net
ULMFiT
Generative Pretraining
ALBERT
SpanBERT
Recursive Deep Net
Doc2vec
Skip-thought Vector
FastText
Deep Pyramid CNN
BERT
RoBERTa
Realm
word2vec
Glove
Character Embedding
Subword Embedding
Transformer
ELMo
XLNet
ELECTRA
2013
2014
2015
2016
2017
2018
2019
2020

Table 1. Accuracy of deep learning based text classification models on sentiment analysis datasets (in terms of classification accuracy), evaluated on the IMDB, SST, Yelp, and Amazon datasets. Italic indicates the non-deep-learning models.

Method	IMDB	SST-2	Amazon-2	Amazon-5	Yelp-2	Yelp-5
Naive Bayes [43]	-	81.80	-	-	-	-
LDA [214]	67.40	-	-	-	-	-
BoW+SVM [31]	87.80	-	-	-	-	-
tf.Δ idf [215]	88.10	-	-	-	-	-
Char-level CNN [50]	-	-	94.49	59.46	95.12	62.05
Deep Pyramid CNN [49]	-	84.46	96.68	65.82	97.36	69.40
ULMFiT [216]	95.40	-	-	-	97.84	70.02
BLSTM-2DCNN [40]	-	89.50	-	-	-	-
Neural Semantic Encoder [95]	-	89.70	-	-	-	-
BCN+Char+CoVe [217]	91.80	90.30	-	-	-	-
GLUE ELMo baseline [22]	-	90.40	-	-	-	-
BERT ELMo baseline [7]	-	90.40	-	-	-	-
CCCapsNet [76]	-	-	94.96	60.95	96.48	65.85
Virtual adversarial training [173]	94.10	-	-	-	-	-
Block-sparse LSTM [218]	94.99	93.20	-	-	96.73	
BERT-base [7, 154]	95.63	93.50	96.04	61.60	98.08	70.58
BERT-large [7, 154]	95.79	94.9	96.07	62.20	98.19	71.38
ALBERT [147]	-	95.20	-	-	-	-
Multi-Task DNN [23]	83.20	95.60	-	-	-	-
Snorkel MeTaL [219]	-	96.20	-	-	-	-
BERT Finetune + UDA [220]	95.80		96.50	62.88	97.95	62.92
RoBERTa (+additional data) [146]	-	96.40	-	-	-	-
XLNet-Large (ensemble) [156]	96.21	96.80	97.60	67.74	98.45	72.20

Table 2. Accuracy of classification models on news categorization, and topic classification tasks. Italic indicates the non-deep-learning models.

	News Categorization			Topic Classification	
Method	AG News	20NEWS	Sogou News	DBpedia	Ohsumed
Hierarchical Log-bilinear Model [221]	-	-	-	-	52
Text GCN [107]	67.61	86.34	-	-	68.36
Simplfied GCN [108]	-	88.50	-	-	68.50
Char-level CNN [50]	90.49	-	95.12	98.45	-
CCCapsNet [76]	92.39	-	97.25	98.72	-
LEAM [84]	92.45	81.91	-	99.02	58.58
fastText [30]	92.50	-	96.80	98.60	55.70
CapsuleNet B [71]	92.60	-	-	-	-
Deep Pyramid CNN [49]	93.13	-	98.16	99.12	-
ULMFiT [216]	94.99	-	-	99.20	-
L MIXED [174]	95.05	-	-	99.30	-
BERT-large [220]	-	-	-	99.32	-
XLNet [156]	95.51	-	-	99.38	-

Table 3. Performance of classification models on SQuAD question answering datasets. Here, the F1 score measures the average overlap between the prediction and ground truth answer. Italic denotes the non-deep-learning models.

	SQuAD1.1		SQuAD2.0	
Method	EM	F1-score	EM	F1-score
Sliding Window+Dist. [222]	13.00	20.00	-	-
Hand-crafted Features+Logistic Regression [24]	40.40	51.00	-	-
BiDAF + Self Attention + ELMo [4]	78.58	85.83	63.37	66.25
SAN (single model) [137]	76.82	84.39	68.65	71.43
FusionNet++ (ensemble) [223]	78.97	86.01	70.30	72.48
SAN (ensemble) [137]	79.60	86.49	71.31	73.70
BERT (single model) [7]	85.08	91.83	80.00	83.06
BERT-large (ensemble) [7]	87.43	93.16	80.45	83.51
BERT + Multiple-CNN [137]	-	-	84.20	86.76
XL-Net [156]	89.90	95.08	84.64	88.00
SpanBERT [149]	88.83	94.63	71.31	73.70
RoBERTa [146]	-	-	86.82	89.79
ALBERT (single model) [147]	-	-	88.10	90.90
ALBERT (ensemble) [147]	-	-	89.73	92.21
Retro-Reader on ALBERT	-	-	90.11	92.58
ELECTRA+ALBERT+EntitySpanFocus	-	-	90.42	92.79

Table 4. Performance of classification models on the WikiQA datasets.

Method	MAP	MRR
Paragraph vector [32]	0.511	0.516
Neural Variational Inference [166]	0.655	0.674
Attentive pooling networks [83]	0.688	0.695
HyperQA [127]	0.712	0.727
BERT (single model) [7]	0.813	0.828
TANDA-RoBERTa [153]	0.920	0.933