Argumentation mining

Paolo Torroni
paolo.torroni@unibo.it

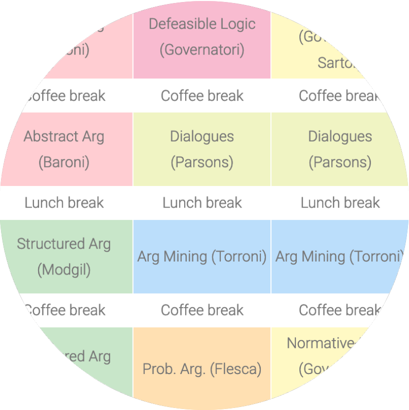
Università degli Studi di Bologna, Italy

Based on joint work with Marco Lippi, Università di Modena e Reggio Emilia; Andrea Galassi and Federico Ruggeri, Università di Bologna

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Introduction

Argumentation in the COMMA world

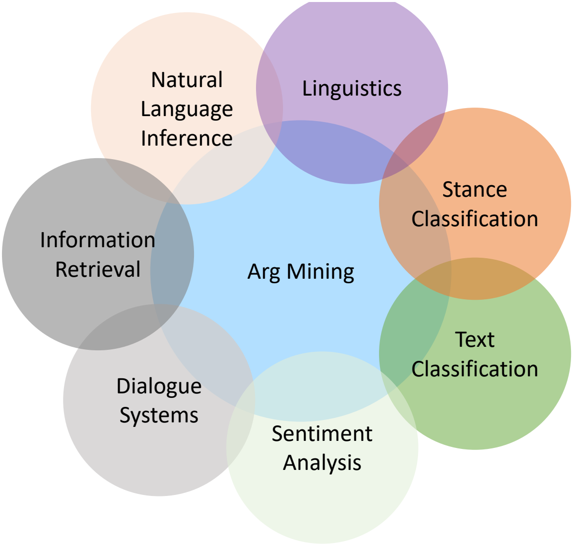


oni)	Defeasible Logic (Governatori)	(Go Sarto
offee break	Coffee break	Coffee brea
Abstract Arg (Baroni)	Dialogues (Parsons)	Dialogues (Parsons)
Lunch break	Lunch break	Lunch break
Structured Arg (Modgil)	Arg Mining (Torroni)	Arg Mining (Torroni
offee break	Coffee break	Coffee break
red Arg	Prob. Arg. (Flesca)	Normative (Gov

Introduction

What do you expect from Argument Mining?

Argumentation in the NLP universe



Why argument mining?

Many possible applications:

- visualization of the main pro and con arguments in a text corpus towards a topic or query of interest
- information management for researchers
- instructional contexts: automated essay grading, critical thinking
- conversational search
- argument search engines
- debating technologies
- social media mining
- public consultations, participatory governance

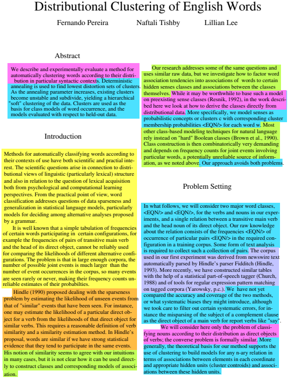
Natural arguments: where to find them?

An incomplete list:

- legal documents
- news articles
- user-generated web discourse
  - Wikipedia articles
  - product reviews
  - online debates, comments, tweets
  - ...
- academic literature
- persuasive essays
- political speech, parliamentary/election debates
- dialogues
- ...

Early days

A first idea: argumentative zoning



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster controids) and associations between these hidden units.

- focus: rhetorical status of sentence with respect to communicative function of the whole paper

S Teufel. Agumentative Zoning: Information Extraction from Scientific Text, PhD Dissertation, U Edinburgh, 1999

Early days

Zones:

- General scientific background
- Neutral descriptions of other people's work
- Neutral descriptions of the own, new work
- Statements of the particular aim of the current paper
- Statements of textual organization of the current paper (in chapter 1, we introduce...)
- Contrastive or comparative statements about other work; explicit mention of weaknesses of other work
- Statements that own work is based on other work

Early days

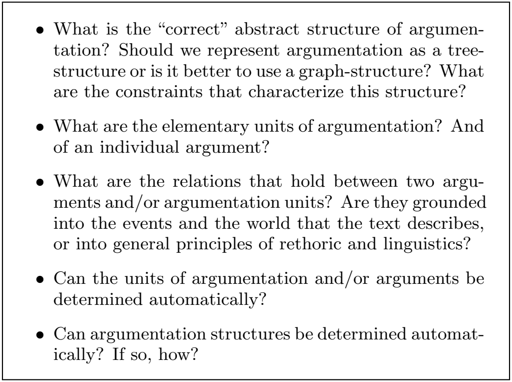
Methodology

- put together corpus
- define pool of sentential features that correlate to a sentence's rhetorical status
- define methods to run analysis based on automatically extracted features
  - statistical classifiers
  - rule-based methods
- evaluate against human performance

A leap forward in time

Argumentation mining: the detection, classification and structure of arguments in text

- Landmark paper by Palau and Moens (ICAIL'09)
- Focus on legal texts; Auracaria and ECHR corpora
- Pinpoints fundamental questions

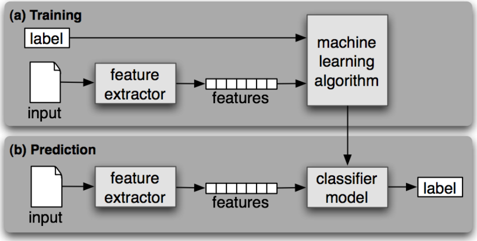


- What is the "correct" abstract structure of argumentation? Should we represent argumentation as a tree-structure or is it better to use a graph-structure? What are the constraints that characterize this structure?
- What are the elementary units of argumentation? And of an individual argument?
- What are the relations that hold between two arguments and/or argumentation units? Are they grounded into the events and the world that the text describes, or into general principles of rethoric and linguistics?
- Can the units of argumentation and/or arguments be determined automatically?
- Can argumentation structures be determined automatically? If so, how?

Experiments

Two different tasks and approaches

- Argument detection/classification: statistical NLP (68% F1)
  - Supervised classifiers, including Naive Bayes and SVM



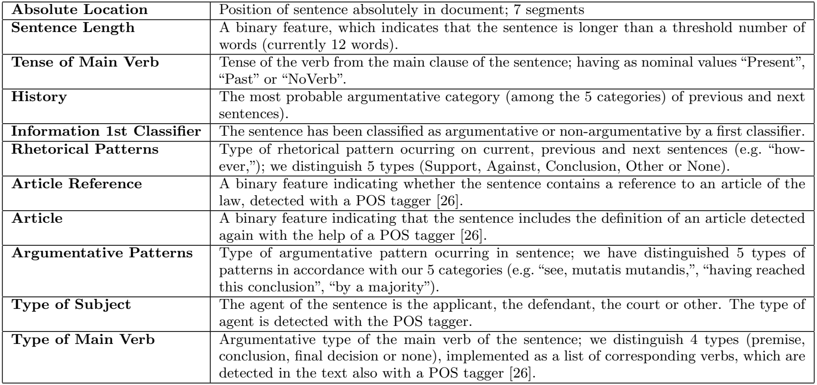
Natural Language Processing with Python, `https://www.nltk.org/book/`

Experiments

Two different tasks and approaches

- Argument detection/classification: statistical NLP (68% F1)
  - Supervised classifiers, including Naive Bayes and SVM

Table 5: Features for the classification of argumentative propositions



Absolute Location	Position of sentence absolutely in document; 7 segments
Sentence Length	A binary feature, which indicates that the sentence is longer than a threshold number of words (currently 12 words).
Tense of Main Verb	Tense of the verb from the main clause of the sentence; having as nominal values "Present", "Past" or "NoVerb".
History	The most probable argumentative category (among the 5 categories) of previous and next sentences).
Information 1st Classifier	The sentence has been classified as argumentative or non-argumentative by a first classifier.
Rhetorical Patterns	Type of rhetorical pattern ocurring on current, previous and next sentences (e.g. "however,"); we distinguish 5 types (Support, Against, Conclusion, Other or None).
Article Reference	A binary feature indicating whether the sentence contains a reference to an article of the law, detected with a POS tagger [26].
Article	A binary feature indicating that the sentence includes the definition of an article detected again with the help of a POS tagger [26].
Argumentative Patterns	Type of argumentative pattern ocurring in sentence; we have distinguished 5 types of patterns in accordance with our 5 categories (e.g. "see, mutatis mutandis,", "having reached this conclusion", "by a majority").
Type of Subject	The agent of the sentence is the applicant, the defendant, the court or other. The type of agent is detected with the POS tagger.
Type of Main Verb	Argumentative type of the main verb of the sentence; we distinguish 4 types (premise, conclusion, final decision or none), implemented as a list of corresponding verbs, which are detected in the text also with a POS tagger [26].

R Mochales Palau & M-F Moens. Argumentation mining: the detection, classification and structure of arguments in text, ICAIL 2009

Experiments

Two different tasks and approaches

- Argument detection/classification: statistical NLP (68% F1)
  - Supervised classifiers, including Naive Bayes and SVM
- Argumentation structure prediction: CFG parsing (70% F1)

$$T \Rightarrow A^+D$$

$$A \Rightarrow \{A^+C|A^*CnP^+|Cns|A^*sr_cC|P^+\}$$

$$D \Rightarrow r_cf\{v_cs|.\}^+$$

$$P \Rightarrow \{P_{verbP}|P_{art}|PP_{sup}|PP_{ag}|sP_{sup}|sP_{ag}\}$$

$$P_{verbP} = sv_ps$$

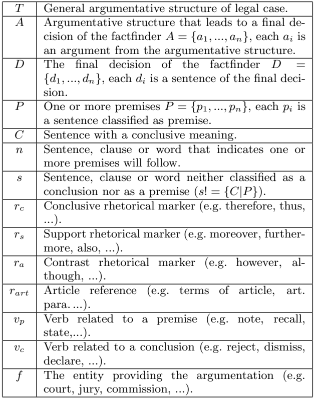
$$P_{art} = sr_{art}s$$

$$P_{sup} = \{r_s\}\{s|P_{verbP}|P_{art}|P_{sup}|P_{ag}\}$$

$$P_{ag} = \{r_a\}\{s|P_{verbP}|P_{art}|P_{sup}|P_{ag}\}$$

$$C = \{r_c|r_s\}\{s|C|r_cP_{verbP}\}$$

$$C = s^*v_cs$$



T	General argumentative structure of legal case.
A	Argumentative structure that leads to a final decision of the factfinder $A = \{a_1, ..., a_n\}$, each a_i is an argument from the argumentative structure.
D	The final decision of the factfinder $D = \{d_1, ..., d_n\}$, each d_i is a sentence of the final decision.
P	One or more premises $P = \{p_1, ..., p_n\}$, each p_i is a sentence classified as premise.
C	Sentence with a conclusive meaning.
n	Sentence, clause or word that indicates one or more premises will follow.
s	Sentence, clause or word neither classified as a conclusion nor as a premise ($s! = \{C \mid P\}$).
r_c	Conclusive rhetorical marker (e.g. therefore, thus, ...).
r_s	Support rhetorical marker (e.g. moreover, furthermore, also, ...).
r_a	Contrast rhetorical marker (e.g. however, although, ...).
r_{art}	Article reference (e.g. terms of article, art. para. ...).
v_p	Verb related to a premise (e.g. note, recall, state,...).
v_c	Verb related to a conclusion (e.g. reject, dismiss, declare, ...).
f	The entity providing the argumentation (e.g. court, jury, commission, ...).

Argument Mining

Main goal:

- **automatically extract arguments** from unstructured text

Emerging potential

- move **sentiment analysis** a step forward and leverage a number of futuristic **applications**
- understand not only opinions, but also **reasons behind them**
- **draw a bridge** between formal models and theories and natural argumentation
- **provide data** for formal argumentation systems

A new area is born

- E Cabrio & S Villata, *Combining textual entailment and argumentation theory for supporting online debates interactions*, ACL 2012
- A Peldszus & M Stede, *From argument diagrams to argumentation mining in texts: A survey*. IJCINI 2013.
- C Stab, I Gurevych, *Identifying Argumentative Discourse Structures in Persuasive Essays*, EMNLP 2014
- 2014: First ArgMining Workshop
    - biomedical texts, essay scoring, user-generated content such as online comments, discussions and short texts
    - Project Debater (IBM) releases first large AM corpus

IBM's Debater

CNET, Jun 19, 2018

Recent years

Rapid evolution of NLP methodologies

- 1980s-1990s: symbolic NLP: rules, CFG
- 1990s-2010s: statistical NPL
- more recently: neural NLP

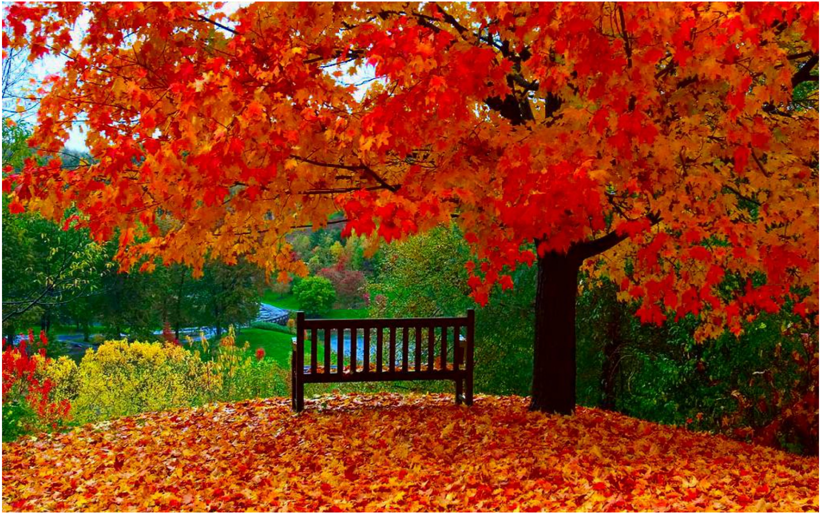
What has changed?

- greater variety of more powerful ML architectures
- less focus on feature engineering
- hunger for large corpora
- tasks have evolved and diversified
- from pipelined to end-to-end systems

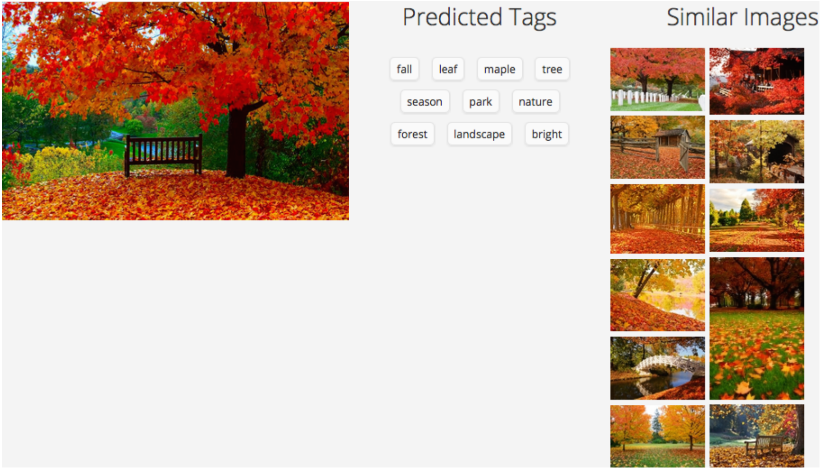
What has not changed?

- *“what is an argument?”*

What do you see in this picture?



Predicted tags and similar images



What is an argument?

“I love bananas”

Example from C Reed & K Budzynska, ACL 2019 tutorial

What is an argument?

“I love bananas”

- “What fruits do you like?” “I love bananas”
- “We should visit the Philippines. I love bananas and they grow amazing ones there - best in the world.”
- “You hate all fruits!” “I love bananas”

Example from C Reed & K Budzynska, ACL 2019 tutorial

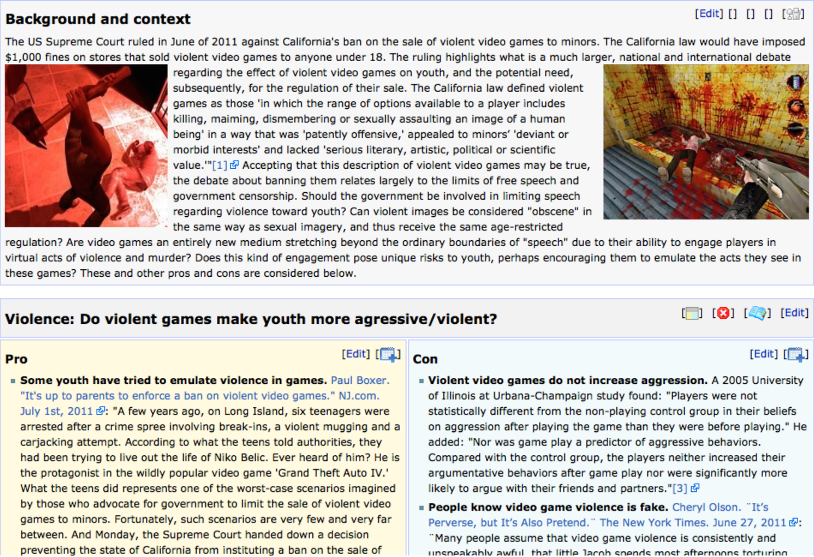
What is an argument?

“I love bananas”

- “What fruits do you like?” “I love bananas” **Not argument**
- “We should visit the Philippines. I love bananas and they grow amazing ones there - best in the world.” **Support**
- “You hate all fruits!” “I love bananas” **Conflict**

Example from C Reed & K Budzynska, ACL 2019 tutorial

Debatepedia: Ban on sale of violent video games to minors



Background and context

[Edit] [] [] [] []

The US Supreme Court ruled in June of 2011 against California's ban on the sale of violent video games to minors. The California law would have imposed $1,000 fines on stores that sold violent video games to anyone under 18. The ruling highlights what is a much larger, national and international debate regarding the effect of violent video games on youth, and the potential need, subsequently, for the regulation of their sale. The California law defined violent games as those 'in which the range of options available to a player includes killing, maiming, dismembering or sexually assaulting an image of a human being' in a way that was 'patently offensive,' appealed to minors' 'deviant or morbid interests' and lacked 'serious literary, artistic, political or scientific value.'"[1] Accepting that this description of violent video games may be true, the debate about banning them relates largely to the limits of free speech and government censorship. Should the government be involved in limiting speech regarding violence toward youth? Can violent images be considered "obscene" in the same way as sexual imagery, and thus receive the same age-restricted regulation? Are video games an entirely new medium stretching beyond the ordinary boundaries of "speech" due to their ability to engage players in virtual acts of violence and murder? Does this kind of engagement pose unique risks to youth, perhaps encouraging them to emulate the acts they see in these games? These and other pros and cons are considered below.

Violence: Do violent games make youth more agressive/violent?

[] [] [] [Edit]

Pro

[Edit] []

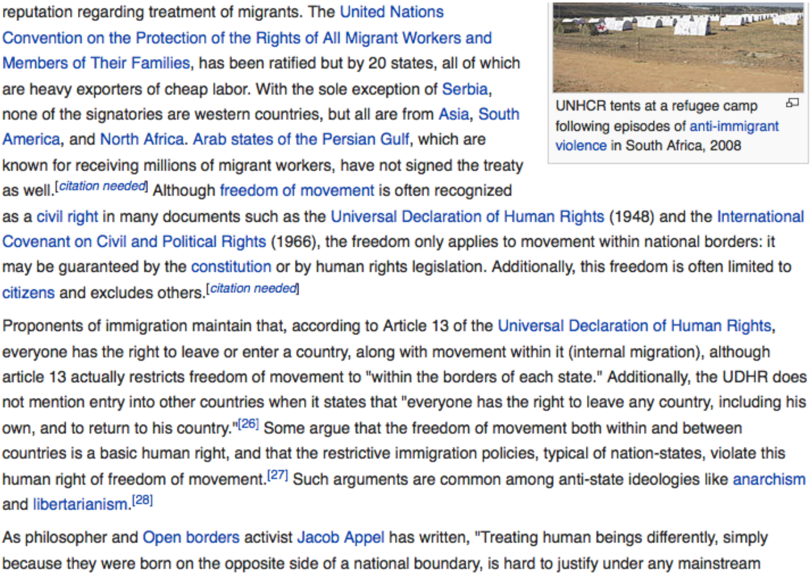
- **Some youth have tried to emulate violence in games.** Paul Boxer. "It's up to parents to enforce a ban on violent video games." NJ.com. July 1st, 2011: "A few years ago, on Long Island, six teenagers were arrested after a crime spree involving break-ins, a violent mugging and a carjacking attempt. According to what the teens told authorities, they had been trying to live out the life of Niko Belic. Ever heard of him? He is the protagonist in the wildly popular video game 'Grand Theft Auto IV.' What the teens did represents one of the worst-case scenarios imagined by those who advocate for government to limit the sale of violent video games to minors. Fortunately, such scenarios are very few and very far between. And Monday, the Supreme Court handed down a decision preventing the state of California from instituting a ban on the sale of

Con

[Edit] []

- **Violent video games do not increase aggression.** A 2005 University of Illinois at Urbana-Champaign study found: "Players were not statistically different from the non-playing control group in their beliefs on aggression after playing the game than they were before playing." He added: "Nor was game play a predictor of aggressive behaviors. Compared with the control group, the players neither increased their argumentative behaviors after game play nor were significantly more likely to argue with their friends and partners."[3]
- **People know video game violence is fake.** Cheryl Olson. ¨It's Perverse, but It's Also Pretend.¨ The New York Times. June 27, 2011: ¨Many people assume that video game violence is consistently and unspeakably awful, that little Jacob spends most afternoons torturing

Wikipedia on Immigration



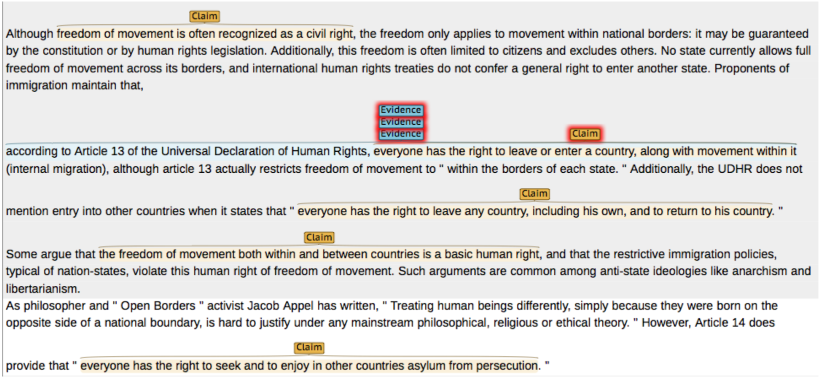
reputation regarding treatment of migrants. The United Nations Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families, has been ratified but by 20 states, all of which are heavy exporters of cheap labor. With the sole exception of Serbia, none of the signatories are western countries, but all are from Asia, South America, and North Africa. Arab states of the Persian Gulf, which are known for receiving millions of migrant workers, have not signed the treaty as well.[citation needed] Although freedom of movement is often recognized as a civil right in many documents such as the Universal Declaration of Human Rights (1948) and the International Covenant on Civil and Political Rights (1966), the freedom only applies to movement within national borders: it may be guaranteed by the constitution or by human rights legislation. Additionally, this freedom is often limited to citizens and excludes others.[citation needed]

UNHCR tents at a refugee camp following episodes of anti-immigrant violence in South Africa, 2008

Proponents of immigration maintain that, according to Article 13 of the Universal Declaration of Human Rights, everyone has the right to leave or enter a country, along with movement within it (internal migration), although article 13 actually restricts freedom of movement to "within the borders of each state." Additionally, the UDHR does not mention entry into other countries when it states that "everyone has the right to leave any country, including his own, and to return to his country."[26] Some argue that the freedom of movement both within and between countries is a basic human right, and that the restrictive immigration policies, typical of nation-states, violate this human right of freedom of movement.[27] Such arguments are common among anti-state ideologies like anarchism and libertarianism.[28]

As philosopher and Open borders activist Jacob Appel has written, "Treating human beings differently, simply because they were born on the opposite side of a national boundary, is hard to justify under any mainstream

Annotations in the IBM Dataset 2014



Although freedom of movement is often recognized as a civil right, the freedom only applies to movement within national borders: it may be guaranteed by the constitution or by human rights legislation. Additionally, this freedom is often limited to citizens and excludes others. No state currently allows full freedom of movement across its borders, and international human rights treaties do not confer a general right to enter another state. Proponents of immigration maintain that,

according to Article 13 of the Universal Declaration of Human Rights, everyone has the right to leave or enter a country, along with movement within it (internal migration), although article 13 actually restricts freedom of movement to " within the borders of each state. " Additionally, the UDHR does not

mention entry into other countries when it states that " everyone has the right to leave any country, including his own, and to return to his country. "

Some argue that the freedom of movement both within and between countries is a basic human right, and that the restrictive immigration policies, typical of nation-states, violate this human right of freedom of movement. Such arguments are common among anti-state ideologies like anarchism and libertarianism.

As philosopher and " Open Borders " activist Jacob Appel has written, " Treating human beings differently, simply because they were born on the opposite side of a national boundary, is hard to justify under any mainstream philosophical, religious or ethical theory. " However, Article 14 does

provide that " everyone has the right to seek and to enjoy in other countries asylum from persecution. "

E Aharoni et al., *A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics*, ArgMining 2014

Problem formulation

Structured argumentation rather than **abstract argumentation**

There is **no unique definition** of a structured argument
$\Rightarrow$ a simple **claim-premise** model is very popular

An example from the IBM corpus

CLAIM
Health risks can be produced by long-term use or excessive doses of anabolic steroids

SUPPORTED BY
A recent study has also shown that long term AAS users were more likely to have symptoms of muscle dysmorphia

Tags

- Labels needed for training automatic classifiers
  - dataset made of pairs (x_i, y_i)
- produced by human annotators
- **Inter-Annotator Agreement** (IAA) is a measure of how well two (or more) annotators can make the same annotation decision for a certain category
  - how trustworthy the annotation?
  - how easy to clearly delineate the category?
- some common metrics: Kohen's κ, Krippendorf's α_U, Pearson's r
- measure the annotations' overlap (modulo the chance agreement)

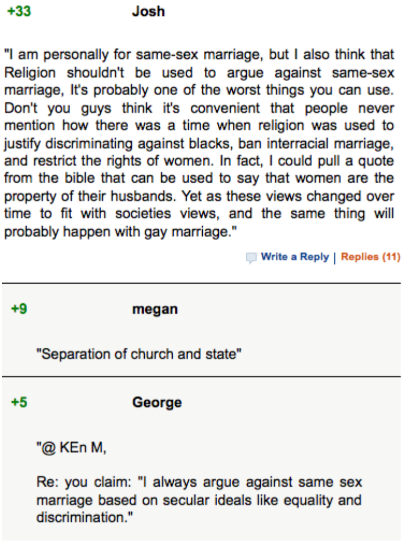
Inter-Annotator Agreement

- Cohen's kappa for classification of N items into C mutually exclusive categories:

  $$\kappa = \frac{p_o - p_e}{1 - p_e}$$

  where p_o is the observed agreement, p_e is chance agreement
  complete agreement: $\kappa = 1$; chance agreement: $\kappa = 0$
- Fleiss' kappa: extension to more than two raters/annotators
- Krippendorf's α_U: similar to above metrics, fixes some issues
- Pearson's r measures linear correlation between variables (-1 $\ldots +1$)
- no "hard" thresholds that make an annotated corpus "bad" or "good" or "good enough"
- indication of best results one can hope for from ML classifier trained on that data
- software libraries

ProCon comments: Should gay marriage be legal?



+33 **Josh**

"I am personally for same-sex marriage, but I also think that Religion shouldn't be used to argue against same-sex marriage, It's probably one of the worst things you can use. Don't you guys think it's convenient that people never mention how there was a time when religion was used to justify discriminating against blacks, ban interracial marriage, and restrict the rights of women. In fact, I could pull a quote from the bible that can be used to say that women are the property of their husbands. Yet as these views changed over time to fit with societies views, and the same thing will probably happen with gay marriage."

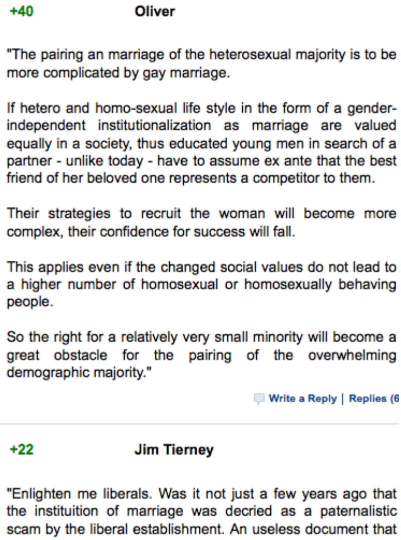
+9 **megan**

"Separation of church and state"

+5 **George**

"@ KEn M,

Re: you claim: "I always argue against same sex marriage based on secular ideals like equality and discrimination."



+40 **Oliver**

"The pairing an marriage of the heterosexual majority is to be more complicated by gay marriage.

If hetero and homo-sexual life style in the form of a gender-independent institutionalization as marriage are valued equally in a society, thus educated young men in search of a partner - unlike today - have to assume ex ante that the best friend of her beloved one represents a competitor to them.

Their strategies to recruit the woman will become more complex, their confidence for success will fall.

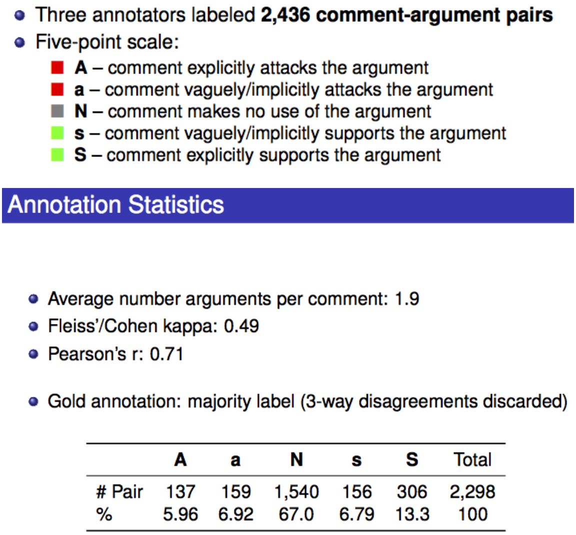
This applies even if the changed social values do not lead to a higher number of homosexual or homosexually behaving people.

So the right for a relatively very small minority will become a great obstacle for the pairing of the overwhelming demographic majority."

+22 **Jim Tierney**

"Enlighten me liberals. Was it not just a few years ago that the instituition of marriage was decried as a paternalistic scam by the liberal establishment. An useless document that

Annotations (Boltužić and Šnajder 2014)



- Three annotators labeled **2,436 comment-argument pairs**
- Five-point scale:
  - 🟥 **A** – comment explicitly attacks the argument
  - 🟥 **a** – comment vaguely/implicitly attacks the argument
  - ◼ **N** – comment makes no use of the argument
  - 🟩 **s** – comment vaguely/implicitly supports the argument
  - 🟩 **S** – comment explicitly supports the argument

Annotation Statistics

- Average number arguments per comment: 1.9
- Fleiss'/Cohen kappa: 0.49
- Pearson's r: 0.71

- Gold annotation: majority label (3-way disagreements discarded)

	A	a	N	s	S	Total
# Pair	137	159	1,540	156	306	2,298
%	5.96	6.92	67.0	6.79	13.3	100

Persuasive essays (UKP)

> *Museums and art galleries provide a better understanding about arts than Internet. In most museums and art galleries, detailed descriptions in terms of the background, history and author are provided. Seeing an artwork online is not the same as watching it with our own eyes, as the picture on line does not show the texture or three-dimensional structure of the art, which is important to study.*

Annotation agreement: $\alpha_U = 0.72$ for argument components, 0.81 for argumentative relations.

C Stab & I Gurevych, *Identifying Argumentative Discourse Structuresin Persuasive Essays*, EMNLP 2014

Argument model in the UKP dataset

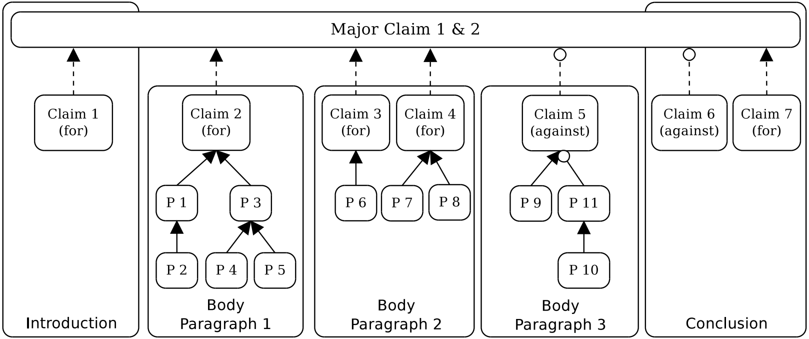
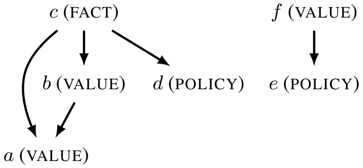


Figure 2
Argumentation structure of the example essay. Arrows indicate argumentative relations. Arrowheads denote argumentative support relations and circleheads attack relations. Dashed lines indicate relations that are encoded in the stance attributes of claims. "P" denotes premises.

C Stab and I Gurevych, *Parsing argumentation structures in persuasive essays*, Computational Linguistics, 2017

Argument model in the CDCP dataset

[Calling a debtor at work is counter-intuitive;]$_a$ [if collectors are continuously calling someone at work, other employees may report it to the debtor's supervisor.]$_b$ [Most companies have established rules about receiving or making personal calls during working hours.]$_c$ [If a collector or creditor calls a debtor on his/her cell phone and is informed that the debtor is at work, the call should be terminated.]$_d$ [No calls to employers should be allowed,]$_e$ [as this jeopardizes the debtor's job.]$_f$



V Niculae et al, *Argument Mining with Structured SVMs and RNNs*, ACL 2017

Annotation agreement

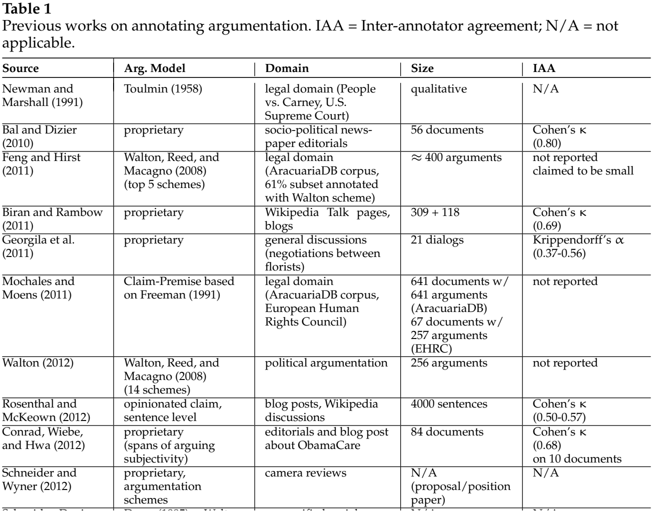


Table 1
Previous works on annotating argumentation. IAA = Inter-annotator agreement; N/A = not applicable.

Source	Arg. Model	Domain	Size	IAA
Newman and Marshall (1991)	Toulmin (1958)	legal domain (People vs. Carney, U.S. Supreme Court)	qualitative	N/A
Bal and Dizier (2010)	proprietary	socio-political newspaper editorials	56 documents	Cohen's κ (0.80)
Feng and Hirst (2011)	Walton, Reed, and Macagno (2008) (top 5 schemes)	legal domain (AracuariaDB corpus, 61% subset annotated with Walton scheme)	≈ 400 arguments	not reported claimed to be small
Biran and Rambow (2011)	proprietary	Wikipedia Talk pages, blogs	309 + 118	Cohen's κ (0.69)
Georgila et al. (2011)	proprietary	general discussions (negotiations between florists)	21 dialogs	Krippendorff's α (0.37-0.56)
Mochales and Moens (2011)	Claim-Premise based on Freeman (1991)	legal domain (AracuariaDB corpus, European Human Rights Council)	641 documents w/ 641 arguments (AracuariaDB) 67 documents w/ 257 arguments (EHRC)	not reported
Walton (2012)	Walton, Reed, and Macagno (2008) (14 schemes)	political argumentation	256 arguments	not reported
Rosenthal and McKeown (2012)	opinionated claim, sentence level	blog posts, Wikipedia discussions	4000 sentences	Cohen's κ (0.50-0.57)
Conrad, Wiebe, and Hwa (2012)	proprietary (spans of arguing subjectivity)	editorials and blog post about ObamaCare	84 documents	Cohen's κ (0.68) on 10 documents
Schneider and Wyner (2012)	proprietary, argumentation schemes	camera reviews	N/A (proposal/position paper)	N/A

Excerpt from C Stab & I Gurevych, *Argumentation Mining in User-Generated Web Discourse*, COLI 2017

Other argument models

- Toulmin's
- Walton's argument schemes
- Inference Anchoring Theory
- Models tailored to specific datasets/genres, e.g., legal texts

Issues with more expressive models

- increased cost of annotation
- large portions of argument not in text, e.g., left implicit
- could be hard to apply even for expert annotators, yielding low IAA
- several studies on this

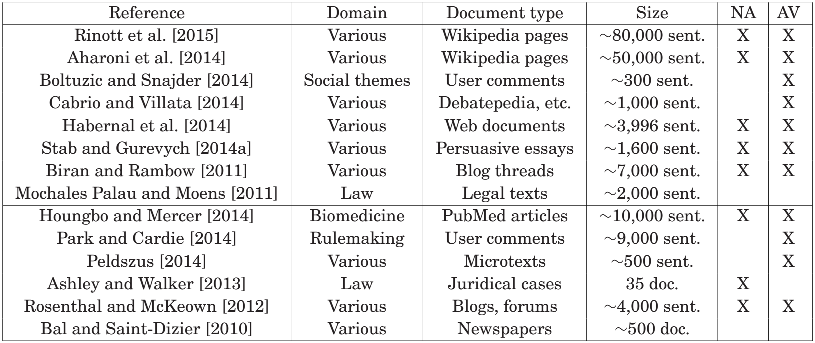
Many attempts at crowdsourcing

- early attempts not very successful, now improving
- trend: "blending" strong/weak annotations

A Lindahl et al, *Towards Assessing Argumentation Annotation – A First Step*, ArgMining 2019
E Musi et al, *A Multi-layer Annotated Corpus of Argumentative Text: From Argument Schemes to Discourse Relations*, ArgMining 2018
T Miller et al, *A Streamlined Method for Sourcing Discourse-level Argumentation Annotations from the Crowd*, NAACL 2019
E Schnarch et al, *Will it Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining*, ACL 2018

Corpora (from Lippi & Torroni 2016)

Table III. English Language Corpora for which There has been Documented Use by AM Systems (Top) or Related Applications (Bottom). For Each Corpus, we Indicate the Domain and Document Type, the Overall Size, whether it Contains Also Nonargumentative Sentences (NA) and whether, at the Time of Writing, they are Publicly Available or Available Upon Request (AV)

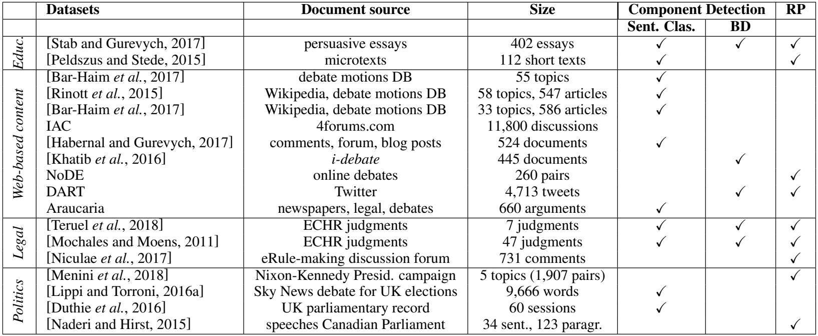


Reference	Domain	Document type	Size	NA	AV
Rinott et al. [2015]	Various	Wikipedia pages	~80,000 sent.	X	X
Aharoni et al. [2014]	Various	Wikipedia pages	~50,000 sent.	X	X
Boltuzic and Snajder [2014]	Social themes	User comments	~300 sent.		X
Cabrio and Villata [2014]	Various	Debatepedia, etc.	~1,000 sent.		X
Habernal et al. [2014]	Various	Web documents	~3,996 sent.	X	X
Stab and Gurevych [2014a]	Various	Persuasive essays	~1,600 sent.	X	X
Biran and Rambow [2011]	Various	Blog threads	~7,000 sent.	X	X
Mochales Palau and Moens [2011]	Law	Legal texts	~2,000 sent.		
Houngbo and Mercer [2014]	Biomedicine	PubMed articles	~10,000 sent.	X	X
Park and Cardie [2014]	Rulemaking	User comments	~9,000 sent.		X
Peldszus [2014]	Various	Microtexts	~500 sent.		X
Ashley and Walker [2013]	Law	Juridical cases	35 doc.	X	
Rosenthal and McKeown [2012]	Various	Blogs, forums	~4,000 sent.	X	X
Bal and Saint-Dizier [2010]	Various	Newspapers	~500 doc.		

See also `http://argumentationmining.disi.unibo.it/resources.html`

M Lippi & P Torroni, *Argumentation Mining: State of the Art and Emerging Trends*, ACM TOIT 2016

Corpora (from Cabrio & Villata 2018)

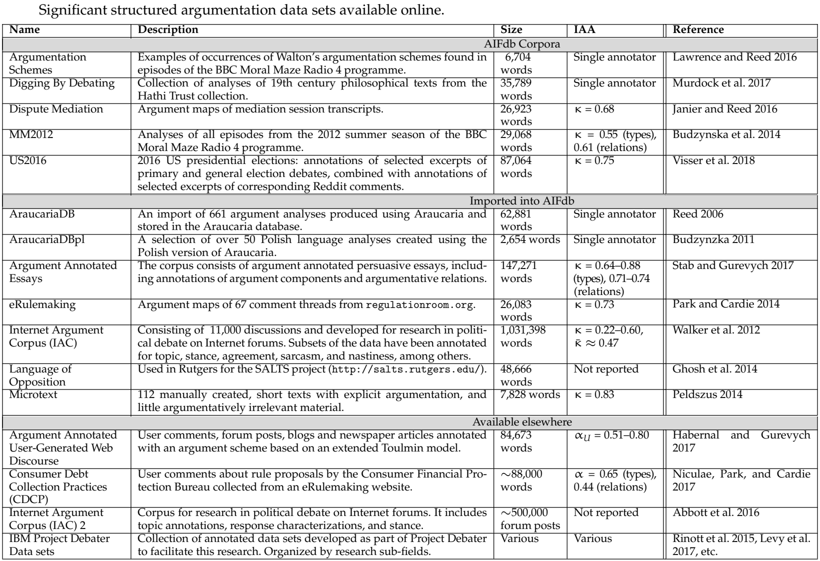


	Datasets	Document source	Size	Component Detection		RP
				Sent. Clas.	BD	
Educ.	[Stab and Gurevych, 2017]	persuasive essays	402 essays	✓	✓	✓
	[Peldszus and Stede, 2015]	microtexts	112 short texts	✓		✓
Web-based content	[Bar-Haim *et al.*, 2017]	debate motions DB	55 topics	✓		
	[Rinott *et al.*, 2015]	Wikipedia, debate motions DB	58 topics, 547 articles	✓		
	[Bar-Haim *et al.*, 2017]	Wikipedia, debate motions DB	33 topics, 586 articles	✓		
	IAC	4forums.com	11,800 discussions			
	[Habernal and Gurevych, 2017]	comments, forum, blog posts	524 documents	✓		
	[Khatib *et al.*, 2016]	*i-debate*	445 documents		✓	
	NoDE	online debates	260 pairs			✓
	DART	Twitter	4,713 tweets		✓	✓
	Araucaria	newspapers, legal, debates	660 arguments	✓		
Legal	[Teruel *et al.*, 2018]	ECHR judgments	7 judgments	✓	✓	✓
	[Mochales and Moens, 2011]	ECHR judgments	47 judgments	✓	✓	✓
	[Niculae *et al.*, 2017]	eRule-making discussion forum	731 comments			✓
Politics	[Menini *et al.*, 2018]	Nixon-Kennedy Presid. campaign	5 topics (1,907 pairs)			✓
	[Lippi and Torroni, 2016a]	Sky News debate for UK elections	9,666 words	✓		
	[Duthie *et al.*, 2016]	UK parliamentary record	60 sessions	✓		
	[Naderi and Hirst, 2015]	speeches Canadian Parliament	34 sent., 123 paragr.			✓

Table 3: Available datasets for AM (sub-)tasks, grouped by their application scenario (BD=boundaries detection; RP=relation prediction).

E Cabrio & S Villata, *Five years of argument mining: A data-driven analysis*, IJCAI, 2018

Corpora (from Lawrence & Reed 2019)



Significant structured argumentation data sets available online.

Name	Description	Size	IAA	Reference
		AIFdb Corpora		
Argumentation Schemes	Examples of occurrences of Walton's argumentation schemes found in episodes of the BBC Moral Maze Radio 4 programme.	6,704 words	Single annotator	Lawrence and Reed 2016
Digging By Debating	Collection of analyses of 19th century philosophical texts from the Hathi Trust collection.	35,789 words	Single annotator	Murdock et al. 2017
Dispute Mediation	Argument maps of mediation session transcripts.	26,923 words	$\kappa = 0.68$	Janier and Reed 2016
MM2012	Analyses of all episodes from the 2012 summer season of the BBC Moral Maze Radio 4 programme.	29,068 words	$\kappa = 0.55$ (types), 0.61 (relations)	Budzynska et al. 2014
US2016	2016 US presidential elections: annotations of selected excerpts of primary and general election debates, combined with annotations of selected excerpts of corresponding Reddit comments.	87,064 words	$\kappa = 0.75$	Visser et al. 2018
		Imported into AIFdb		
AraucariaDB	An import of 661 argument analyses produced using Araucaria and stored in the Araucaria database.	62,881 words	Single annotator	Reed 2006
AraucariaDBpl	A selection of over 50 Polish language analyses created using the Polish version of Araucaria.	2,654 words	Single annotator	Budzynzka 2011
Argument Annotated Essays	The corpus consists of argument annotated persuasive essays, including annotations of argument components and argumentative relations.	147,271 words	$\kappa = 0.64$–0.88 (types), 0.71–0.74 (relations)	Stab and Gurevych 2017
eRulemaking	Argument maps of 67 comment threads from `regulationroom.org`.	26,083 words	$\kappa = 0.73$	Park and Cardie 2014
Internet Argument Corpus (IAC)	Consisting of 11,000 discussions and developed for research in political debate on Internet forums. Subsets of the data have been annotated for topic, stance, agreement, sarcasm, and nastiness, among others.	1,031,398 words	$\kappa = 0.22$–0.60, $\bar{\kappa} \approx 0.47$	Walker et al. 2012
Language of Opposition	Used in Rutgers for the SALTS project (`http://salts.rutgers.edu/`).	48,666 words	Not reported	Ghosh et al. 2014
Microtext	112 manually created, short texts with explicit argumentation, and little argumentatively irrelevant material.	7,828 words	$\kappa = 0.83$	Peldszus 2014
		Available elsewhere		
Argument Annotated User-Generated Web Discourse	User comments, forum posts, blogs and newspaper articles annotated with an argument scheme based on an extended Toulmin model.	84,673 words	$\alpha_U = 0.51$–0.80	Habernal and Gurevych 2017
Consumer Debt Collection Practices (CDCP)	User comments about rule proposals by the Consumer Financial Protection Bureau collected from an eRulemaking website.	∼88,000 words	$\alpha = 0.65$ (types), 0.44 (relations)	Niculae, Park, and Cardie 2017
Internet Argument Corpus (IAC) 2	Corpus for research in political debate on Internet forums. It includes topic annotations, response characterizations, and stance.	∼500,000 forum posts	Not reported	Abbott et al. 2016
IBM Project Debater Data sets	Collection of annotated data sets developed as part of Project Debater to facilitate this research. Organized by research sub-fields.	Various	Various	Rinott et al. 2015, Levy et al. 2017, etc.

Major AM resource portals

- IBM Debater project datasets (Slonim et al)
  `www.research.ibm.com/haifa/dept/vst/debating_data.shtml`
- UKP Darmstadt (Gurevych et al)
  `www.informatik.tu-darmstadt.de/ukp/research_6/data`
- ARG-Tech corpora Dundee (Reed et al)
  `corpora.aifdb.org`

Corpora: summary

Building corpora for AM is:

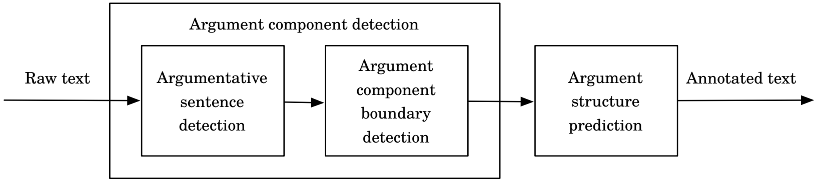
- controversial
- difficult
- time-consuming
- ... necessary ☺

Main limitations of current corpora:

- in most cases **highly domain-dependent**
- sometimes **lacking non-argumentative text**
- generally **adopting custom labels**
- sometimes **very small**, difficult to crowdsource
- need to watch out for **quality of annotations**
- difficult to assess **cross-dataset** performance

The (Old?) Argumentation Mining Pipeline

A typical argument mining problem could be divided conceptually into subsequent subtasks (stages):


Raw text
Argument component detection
Argumentative sentence detection
Argument component boundary detection
Argument structure prediction
Annotated text

Nowadays, there is a growing number of approaches that aim to **jointly** address all these stages.

M Lippi & P Torroni, *Argumentation Mining: State of the Art and Emerging Trends*, ACM TOIT 2016

Sample AM process following the pipeline

CLAIM 1 While those on the far-right think that immigration threatens national identity, as well as cheapening labor and increasing dependence on welfare.

[...]

Proponents of immigration maintain that, according to Article 13 of the Universal Declaration of Human Rights, everyone has the right to leave or enter a country, along with movement within it [...] **EVIDENCE 2**

[...]

CLAIM 3 Some argue that the freedom of movement both within and between countries is a basic human right, and that the restrictive immigration policies, typical of nation-states, violate this human right of freedom of movement.

[...]

Immigration has been a major source of population growth and cultural change throughout much of the history of Sweden. The economic, social, and political aspects of immigration have caused controversy regarding ethnicity, economic benefits, jobs for non-immigrants, settlement patterns, impact on upward social mobility, crime, and voting behavior. **EVIDENCE 4**

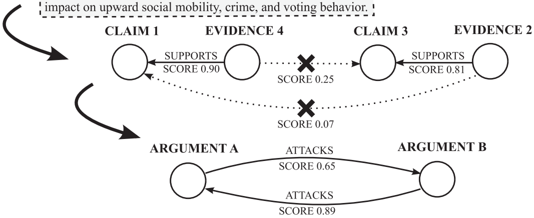


Fig. 1. Example of argument extraction from plain text.

M Lippi & P Torroni, *Argumentation Mining: State of the Art and Emerging Trends*, ACM TOIT 2016

First Stage: Sentence Classification

Predict whether a sentence **is argumentative or not**.

A classic classification task:

- observations $\mathcal{X}$ (sentences)
- labels $\mathcal{Y}$ (e.g., argumentative or not)
- data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$
- find a function $f : \mathcal{X} \rightarrow \mathcal{Y}$
- given a new example $\hat{x} \in \mathcal{X}$, find $\hat{y} = f(\hat{x})$

The task upon which **most (earlier) work** has been spent

Sentence Classification

The labeling is what **defines** the problem:

- distinguish **argumentative** sentences from those that do not contain any argument component
- detect sentences containing **claims**
- detect sentences containing **evidence**
- perform **multi-class** classification
- can be **topic-dependent** or not
- can be **context-dependent** or not

Argument component boundary detection

A segmentation (or **sequence labeling**) problem:

- **second step** of the pipeline, following sentence classification
- needed to detect the **portions** of sentences containing argument components

Different cases can be distinguished:

1. only a portion of the sentence coincides with an argument components;
2. two or more argument components can be present within the same sentence;
3. an argument component can span across multiple sentences.

Argument component boundary detection

Claim example taken from the IBM claim/evidence corpus

A significant number of republicans assert that **hereditary monarchy is unfair and elitist**

Evidence example taken from the IBM claim/evidence corpus

When New Hampshire authorized a state lottery in 1963, it represented a major shift in social policy. No state governments had previously directly run gambling operations to raise money. Other states followed suit, and now the majority of the states run some type of lottery to raise funds for state operations.

Argument component boundary detection

Not many approaches for this task:

- Maximum Likelihood classifiers
- Conditional Random Fields
- Hidden Markov Support Vector Machines (SVM-HMM)

This is a **sequence labeling** task, so **structured-output** or **relational learning** classifiers should be employed !

Given an input sentence $s = \{s_1, ..., s_k\}$
the goal is to produce an output tag sequence $t = \{t_1, ..., t_k\}$

Argument component detection

Traditional machine learning algorithms

- consider all the examples **independently**
- do not take into account **relations** between them

Relational machine learning algorithms

- can exploit **relations** such as data sequentiality
- can produce an overall output through **collective classification**

Argumentation structure prediction

Predict **links** between two argument components (e.g., premise supporting claim) and/or between two arguments (e.g., one argument attacking another one)

- Also this task strongly depends on the **underlying model**
- The **most complex** task in argument mining
- In humans, it typically requires **reasoning** tasks
- Can it be addressed **jointly** to the first stages?

The (new?) argumentation mining steps

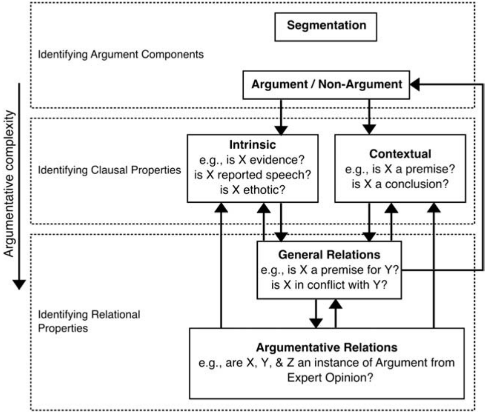


Figure 3
The tasks and levels of complexity in argument mining techniques.

J Lawrence & C Reed, *Argument mining: A survey*, Computational Linguistics, 2019

Techniques

What are the most typical techniques used in AM?

Argument component detection/classification

- **Statistical classifiers** with **handcrafted features** (lexicon, discourse markers, part-of-speech)
- Different instantiations of **deep neural networks** (recurrent, convolutional, attention-based, ...)

Techniques

What are the most typical techniques used in AM?

Predicting the structure of argument graphs

- **Statistical classifiers** with **handcrafted features** (lexicon, discourse markers, part-of-speech)
- Different instantiations of **deep neural networks** (recurrent, convolutional, attention-based, ...)
- **Symbolic approaches** (e.g., textual entailment)
- **Structured output machine learning** methods

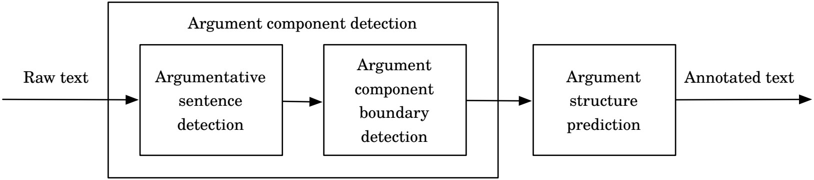
Techniques

Structured output machine learning methods

- **Constraints** amongst argument components
- Example: if a premise supports two claims, then such claims cannot attack each other
- **Many** implementations: ILP, factor graphs, structured SVMs
- Potential in **neural-symbolic** learning or **statistical relational** learning

A Galassi et al: *Neural-Symbolic Argumentation Mining: An Argument in Favor of Deep Learning and Reasoning*, Frontiers in Big Data, 2020

Claim Detection


Argument component detection
Raw text
Argumentative sentence detection
Argument component boundary detection
Argument structure prediction
Annotated text

Claim Detection: predict whether a sentence **contains a claim**

- widely-used claim/premise argument model
- a classic classification task
- basic form of argumentation mining

Sentence Classification

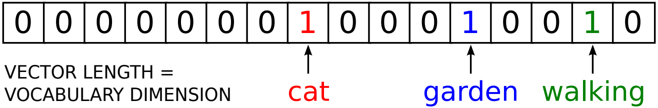
The main point is: **how to encode information about sentences** ?

A classic problem in NLP

- Bag-of-words
- TF-IDF
- Part-of-Speech
- Keyphrases and lists
- Ontologies
- Distributed Representations

Bag-of-Words

`The cat is walking in the garden`



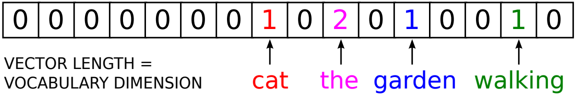
0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0

Represent a sentence with a feature vector and then run any machine learning classifier to discriminate among different classes: features have to **capture similarities** between examples of the same class !

TF-IDF

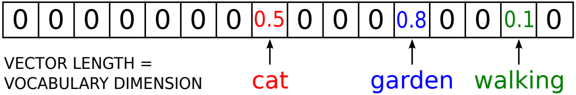
Other BoW variants: consider **frequencies**

- frequency of a word **within a document**
  - Term Frequency (TF)



0	0	0	0	0	0	0	1	0	2	0	1	0	0	1	0

- frequency of a word **within a corpus**:
  - Inverse Document Frequency (IDF)

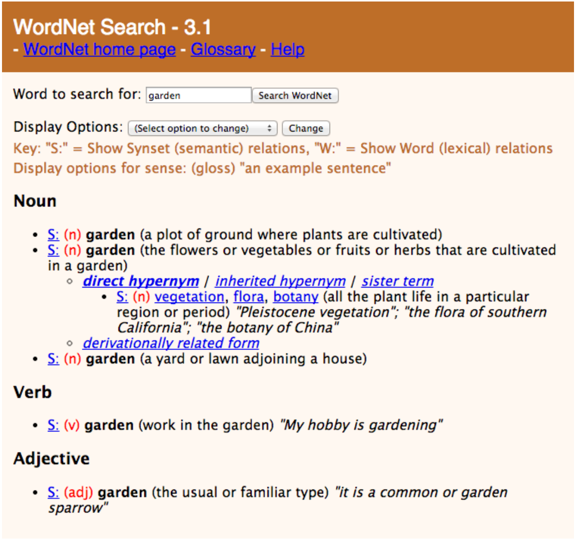


0	0	0	0	0	0	0	0.5	0	0	0	0.8	0	0	0.1	0

Having **rare words** in common is much more significant...
...But still, **it is not enough** !

Ontologies

WordNet is a large ontology or linguistic-semantic database.



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: garden [Search WordNet]

Display Options: (Select option to change) [Change]
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **garden** (a plot of ground where plants are cultivated)
- S: (n) **garden** (the flowers or vegetables or fruits or herbs that are cultivated in a garden)
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) vegetation, flora, botany (all the plant life in a particular region or period) *"Pleistocene vegetation"; "the flora of southern California"; "the botany of China"*
  - *derivationally related form*
- S: (n) **garden** (a yard or lawn adjoining a house)

Verb

- S: (v) **garden** (work in the garden) *"My hobby is gardening"*

Adjective

- S: (adj) **garden** (the usual or familiar type) *"it is a common or garden sparrow"*

Part-of-Speech

An important source of information is also given by **the grammar of a sentence**...

Part-of-speech: word category (noun, verb, adjective, ...)

- try to associate each word to its PoS
- describe each sentence also in terms of PoS
- for example, use a Bag-of-PoS

Part-of-Speech

Part-of-Speech **tagger**: associate a PoS-tag to each word in a sentence.

DT/The NN/cat VBZ/is VBG/walking IN/in DT/the NN/garden

Keyphrases

Sometimes it is useful to use pre-computed lists of **known argumentative phrases or words**:

- use a Bag-of-Keyphrases
- if a keyphrase is found in a sentence, activate a feature

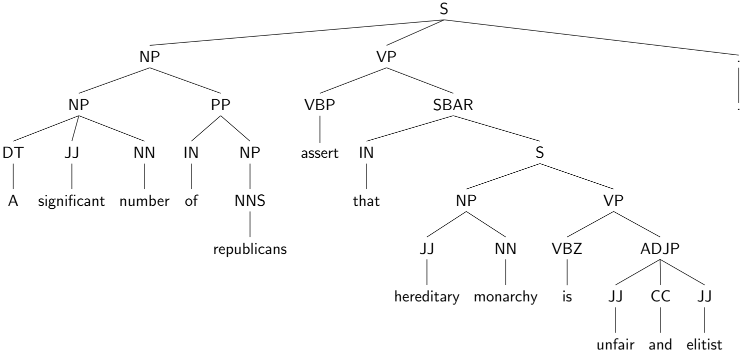
These are powerful features, but...

- they are **hand-tailored**
- necessary to update them by hand
- highly **context-dependent** ?
- do they really **generalize** ?

R Palau & MF Moens, *Argumentation mining*. Artificial Intelligence and Law, 2011
C Stab & I Gurevych, *Identifying argumentative discourse structures in persuasive essays*, EMNLP14
R Levy et al., *Context dependent claim detection*, COLING14
R Rinott et al., *Show Me Your Evidence – an Automatic Method for Context-Dependent Evidence Detection*, EMNLP15

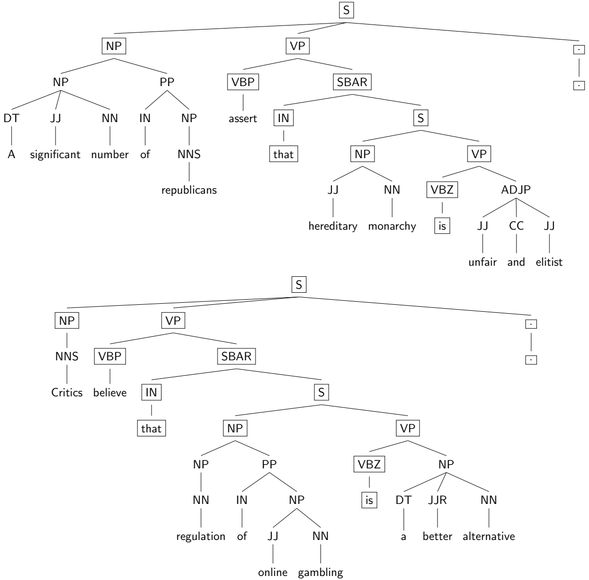
Tree Kernels

Main Idea: the **parse tree** of a sentence is highly indicative of the presence of a claim, as it encodes **rhetorical structure** information



A Moschitti, *Making Tree Kernels Practical for Natural Language Learning*, EACL 2006

Tree Kernels



Context-Independent Claim Detection

Build a **kernel machine** classifier exploiting similarity between trees

- measure similarity between the **structure of sentences**
- count common substructures or **fragments** between trees (Δ)
- we consider the **Partial Tree Kernel** (PTK)

Find function $f : \mathcal{X} \rightarrow \mathcal{Y}$

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x)$$

$$K(T_x, T_z) = \sum_{n_x \in N_{T_x}} \sum_{n_z \in N_{T_z}} \Delta(n_x, n_z)$$

M Lippi & P Torroni, *Context-Independent Claim Detection*, IJCAI 2015

Experiments

IBM Dataset (2014 version)

- around 50,000 sentences from Wikipedia pages
- organized in 33 topics
- around 1,500 annotated **context-dependent** claims

Persuasive essay corpus (Stab & Gurevych, 2014)

- 90 documents (essays)
- around 1,000 sentences
- heterogeneous topics

Qualitative results on 10 additional Wikipedia pages

Experimental results (IBM corpus)

Method	P@200	R@200	F_1@200	AURPC	AURORC
TK	9.8	58.7	16.8	0.161	0.808
BoW	8.2	51.7	14.2	0.117	0.771
Random Baseline	2.8	20.4	5.0	–	–
Perfect Baseline	19.6	99.3	32.7	–	–
TK + Topic	10.5	62.9	18.0	0.178	0.823
IBM Results	9.0	73.0	16.0	–	–

Model is **also** capable of identifying **topic-independent** claims.

Experimental results (IBM corpus)

The data set is context-dependent, thus some of the examples that we predict as claims are actually labeled as negative examples...

IBM Corpus Topic	Sentence
All nations have a right to nuclear weapons	Critics argue that this would lower the threshold for use of nuclear weapons
Atheism is the only way	Some believe that a moral sense does not depend on religious belief
Endangered species should be protected	Simple logic instructs that more people will require more food
Institute a mandatory retirement age	Some theories suggest that ageing is a disease
Limit the right to bear arms	Others doubt that gun controls possess any preventative efficacy
Make physical education compulsory	Specific training prepares athletes to perform well in their sports
Multiculturalism	Indigenous peoples have the right to self-determination

Experimental results

Persuasive essays corpus:

74.6/68.4 precision/recall

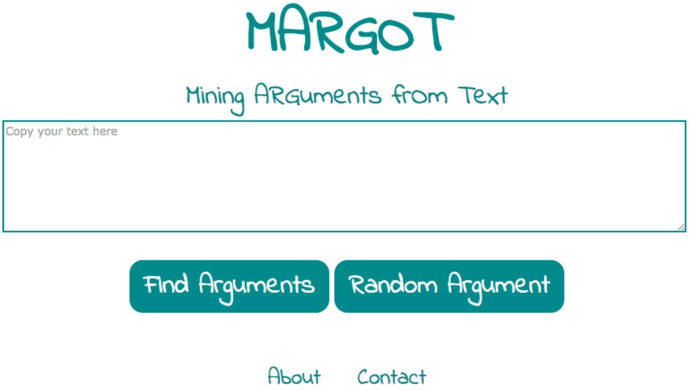
Qualitative results on 10 additional Wikipedia pages:

5 articles on **controversial** topics: Anti-consumerism, Effects of climate change on wine production, Delegative democracy, Geothermal heating, Software patents and free software

5 on **non-controversial** topics: Ethernet, Giardini Naxos, Iamb, Penalty kick, Spacecraft

34 vs. 3 claims detected in the two datasets

A web service for argument mining



`http://margot.disi.unibo.it`

M Lippi & P Torroni, *MARGOT: A Web Server for Argumentation Mining*, Expert Systems with Applications 2016

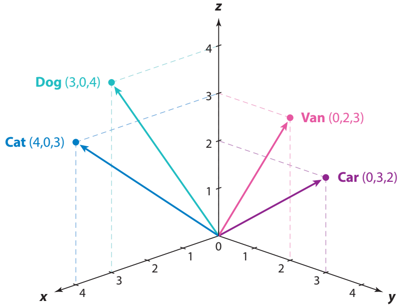
Distributed Representations of Word Meaning

Harris (1954): "Language is not merely a bag of words"
Firth (1957): "[You shall know a word] by the company it keeps"

$$\begin{array}{cccccccccc} & \textit{bite} & \textit{buy} & \textit{drive} & \textit{eat} & \textit{get} & \textit{live} & \textit{park} & \textit{ride} & \textit{tell} \\ \textit{bike} & 0 & 9 & 0 & 0 & 12 & 0 & 8 & 6 & 0 \\ \textit{car} & 0 & 13 & 8 & 0 & 15 & 0 & 5 & 0 & 0 \\ \textit{dog} & 0 & 0 & 0 & 9 & 10 & 7 & 0 & 0 & 1 \\ \textit{lion} & 6 & 0 & 0 & 1 & 8 & 3 & 0 & 0 & 0 \end{array}$$

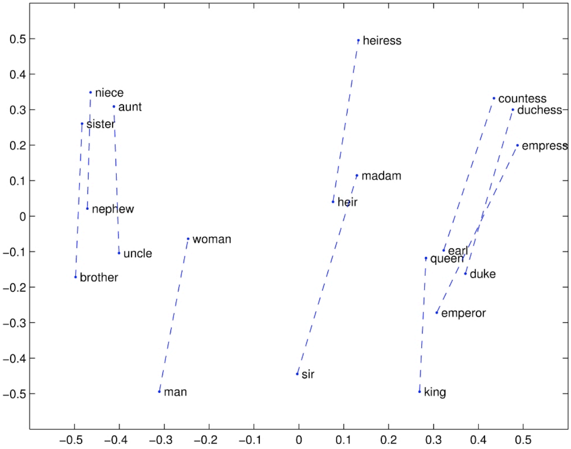
ZS Harris, *Distributional structure*, Word, 1954
JR Firth, *A synopsis of linguistic theory, 1930–1955*, Studies in Linguistic Analysis, 1957
A Lenci, *Distributional Models of Word Meaning*, Annual Review of Linguistics, 2018

Distributed Representations



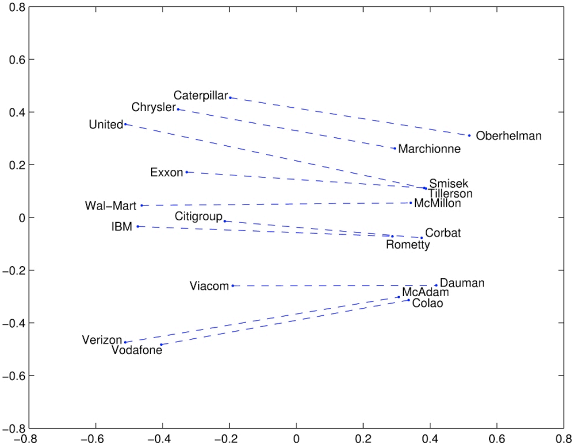
A Lenci, *Distributional Models of Word Meaning*, Annual Review of Linguistics, 2018

GLoVe Visualizations



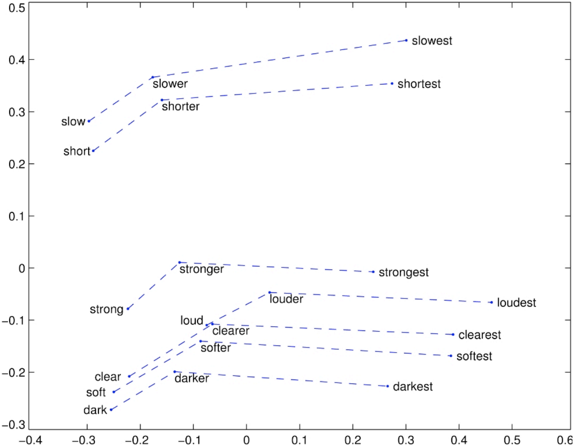
Pictures from `http://web.stanford.edu/class/cs224n/`

GLoVe Visualizations



Pictures from `http://web.stanford.edu/class/cs224n/`

GLoVe Visualizations



Pictures from `http://web.stanford.edu/class/cs224n/`

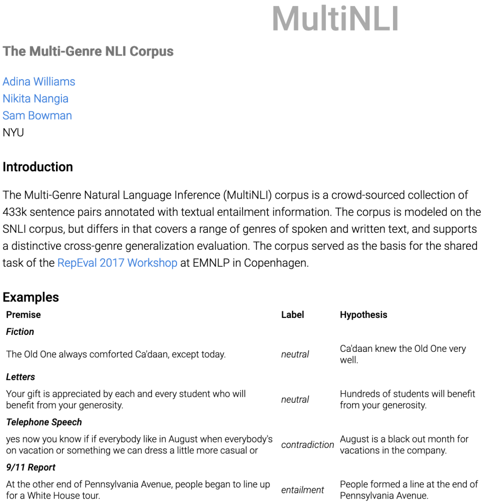
Contextual Word Representations

A stream of widely accepted models

- word2vec (2013)
- GLoVe (2014)
- ELMo (2018)
- BERT (2019)
- GPT-2
- RoBERTa
- T5
- ALBERT
- XLM
- XLNet
- GPT-3
- ...

NA Smith, *Contextual Word Representations: Putting Words into Computers*, CACM, June 2020
For fun: `https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html`

Natural Language Inference



MultiNLI

The Multi-Genre NLI Corpus

Adina Williams
Nikita Nangia
Sam Bowman
NYU

Introduction

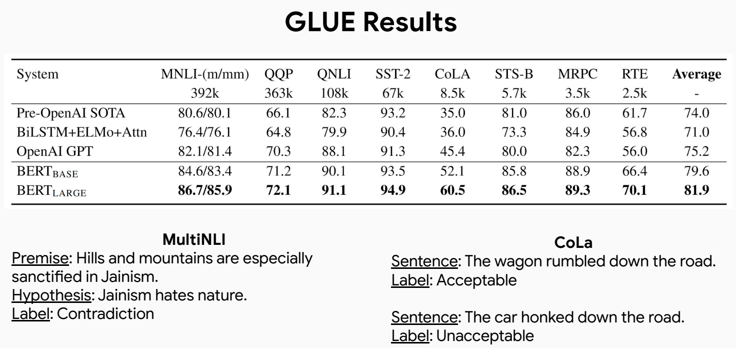
The Multi-Genre Natural Language Inference (MultiNLI) corpus is a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information. The corpus is modeled on the SNLI corpus, but differs in that covers a range of genres of spoken and written text, and supports a distinctive cross-genre generalization evaluation. The corpus served as the basis for the shared task of the RepEval 2017 Workshop at EMNLP in Copenhagen.

Examples

Premise	Label	Hypothesis
Fiction		
The Old One always comforted Ca'daan, except today.	*neutral*	Ca'daan knew the Old One very well.
Letters		
Your gift is appreciated by each and every student who will benefit from your generosity.	*neutral*	Hundreds of students will benefit from your generosity.
Telephone Speech		
yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	*contradiction*	August is a black out month for vacations in the company.
9/11 Report		
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	*entailment*	People formed a line at the end of Pennsylvania Avenue.

https://cims.nyu.edu/~sbowman/multinli/

Contextual Word Vectors: A Game-Changer for NLP ☺



GLUE Results

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT$_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT$_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

MultiNLI

Premise: Hills and mountains are especially sanctified in Jainism.
Hypothesis: Jainism hates nature.
Label: Contradiction

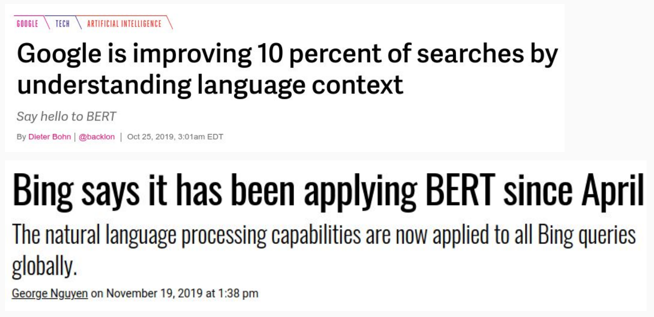
CoLa

Sentence: The wagon rumbled down the road.
Label: Acceptable

Sentence: The car honked down the road.
Label: Unacceptable

Pictures from a talk by J Devlin

Contextual Word Vectors: A Game-Changer for NLP ☺



GOOGLE \ TECH \ ARTIFICIAL INTELLIGENCE

Google is improving 10 percent of searches by understanding language context

Say hello to BERT

By Dieter Bohn | @backlon | Oct 25, 2019, 3:01am EDT

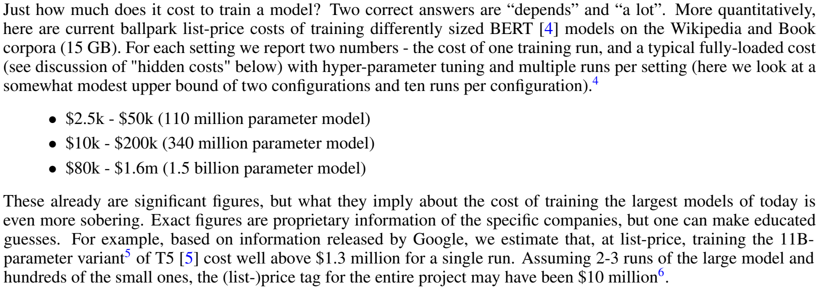
Bing says it has been applying BERT since April

The natural language processing capabilities are now applied to all Bing queries globally.

George Nguyen on November 19, 2019 at 1:38 pm

Pictures from a talk by J Devlin

The Cost of Training NLP Models ☹



Just how much does it cost to train a model? Two correct answers are “depends” and “a lot”. More quantitatively, here are current ballpark list-price costs of training differently sized BERT [4] models on the Wikipedia and Book corpora (15 GB). For each setting we report two numbers - the cost of one training run, and a typical fully-loaded cost (see discussion of "hidden costs" below) with hyper-parameter tuning and multiple runs per setting (here we look at a somewhat modest upper bound of two configurations and ten runs per configuration).[4]

- $2.5k - $50k (110 million parameter model)
- $10k - $200k (340 million parameter model)
- $80k - $1.6m (1.5 billion parameter model)

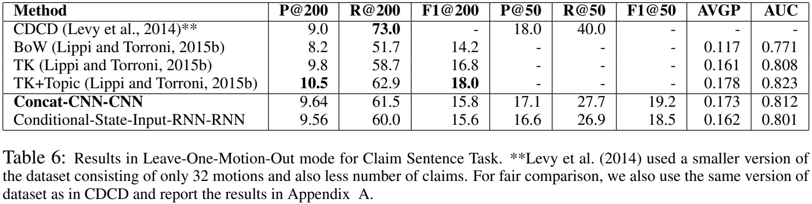
These already are significant figures, but what they imply about the cost of training the largest models of today is even more sobering. Exact figures are proprietary information of the specific companies, but one can make educated guesses. For example, based on information released by Google, we estimate that, at list-price, training the 11B-parameter variant[5] of T5 [5] cost well above $1.3 million for a single run. Assuming 2-3 runs of the large model and hundreds of the small ones, the (list-)price tag for the entire project may have been $10 million[6].

O Sharir et al., *The Cost of Training NLP Models: A Concise Overview*, arXiv:2004.08900, April 2020

Argument Component Detection

Problem: Detecting simple argument components

- Still a fundamental building block of AM systems
- Sometimes good baselines obtained with classic approaches



Method	**P@200**	**R@200**	**F1@200**	**P@50**	**R@50**	**F1@50**	**AVGP**	**AUC**
CDCD (Levy et al., 2014)**	9.0	**73.0**	-	18.0	40.0	-	-	-
BoW (Lippi and Torroni, 2015b)	8.2	51.7	14.2	-	-	-	0.117	0.771
TK (Lippi and Torroni, 2015b)	9.8	58.7	16.8	-	-	-	0.161	0.808
TK+Topic (Lippi and Torroni, 2015b)	**10.5**	62.9	**18.0**	-	-	-	0.178	0.823
Concat-CNN-CNN	9.64	61.5	15.8	17.1	27.7	19.2	0.173	0.812
Conditional-State-Input-RNN-RNN	9.56	60.0	15.6	16.6	26.9	18.5	0.162	0.801

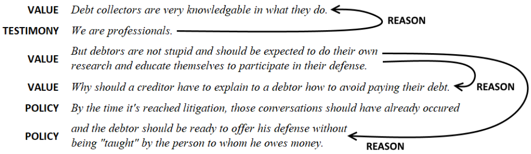
Table 6: Results in Leave-One-Motion-Out mode for Claim Sentence Task. **Levy et al. (2014) used a smaller version of the dataset consisting of only 32 motions and also less number of claims. For fair comparison, we also use the same version of dataset as in CDCD and report the results in Appendix A.

A Laha & V Raykar, *An Empirical Evaluation of various Deep Learning Architectures for Bi-Sequence Classification Tasks*, COLING 2016

Argument Structure Prediction

Problem: inferring relations among arguments

- one of the hardest tasks in AM (implicit context, number of pairs)
- joint learning seems promising
- contextual word embeddings



VALUE *Debt collectors are very knowledgable in what they do.*

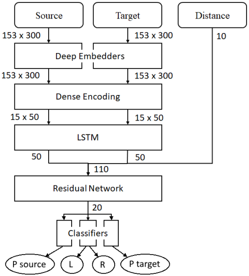
TESTIMONY *We are professionals.*

VALUE *But debtors are not stupid and should be expected to do their own research and educate themselves to participate in their defense.*

VALUE *Why should a creditor have to explain to a debtor how to avoid paying their debt.*

POLICY *By the time it's reached litigation, those conversations should have already occured*

POLICY *and the debtor should be ready to offer his defense without being "taught" by the person to whom he owes money.*

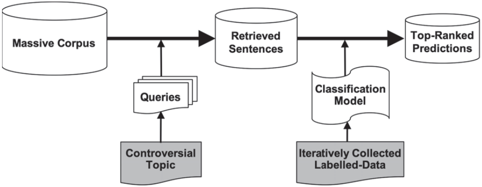


O Cocarascu, F Toni, *Identifying attack and support argumentative relations using deep learning*. EMNLP 2017
HV Nguyen & DJ Litman, *Context-aware Argumentative Relation Mining*, ACL 2016
I Persing & V Ng, *End-to-end argumentation mining in student essays*, NAACL 2016
S Eger et al., *Neural end-to-end learning for computational argumentation mining*, ACL 2017
V Niculae et al, *Argument Mining with Structured SVMs and RNNs*, ALC 2017
A Galassi et al, *Argumentative Link Prediction using Residual Networks and Multi-Objective Learning*, ArgMining 2018
G Morio et al, *Towards Better Non-Tree Argument Mining: Proposition-Level Biaffine Parsing with Task-Specific Parameterization*, ACL 2020

Argument Ranking and Retrieval

Problem: retrieve relevant argument from large corpora and shortlist

- ranking evaluation: absolute? user-dependent?
- argument relevance and quality both matter and are challenging
- need to find counter-arguments

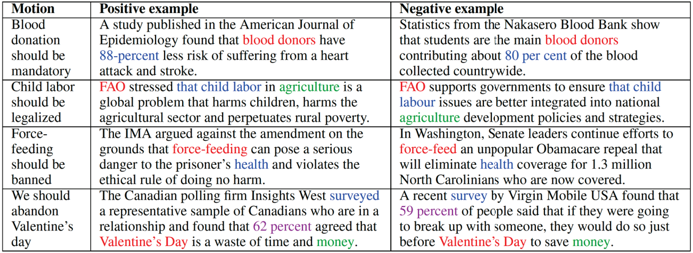


M Lippi et al, *Argumentative Ranking*, NLP Meets Journalism @AAAI 2016
K Al-Khatib et al, *Cross-Domain Mining of Argumentative Text through Distant Supervision*, NAACL 2016
H Wachsmuth et al, *Building an Argument Search Engine for the Web*, ArgMining 2017
M Orbach et al, *Out of the Echo Chamber: Detecting Countering Debate Speeches*, ACL 2020
L Ein-Dor et al, *Corpus Wide Argument Mining—A Working Solution*, AAAI 2020
S Gretz et al., *A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis*, AAAI 2020
JW Sirrianni, *Agreement Prediction of Arguments in Cyber Argumentation for Detecting Stance Polarity and Intensity*, ACL 2020

Argument Ranking and Retrieval

Problem: retrieve relevant argument from large corpora and shortlist

- ranking evaluation: absolute? user-dependent?
- argument relevance and quality both matter and are challenging
- need to find counter-arguments



Motion	Positive example	Negative example
Blood donation should be mandatory	A study published in the American Journal of Epidemiology found that blood donors have 88-percent less risk of suffering from a heart attack and stroke.	Statistics from the Nakasero Blood Bank show that students are the main blood donors contributing about 80 per cent of the blood collected countrywide.
Child labor should be legalized	FAO stressed that child labor in agriculture is a global problem that harms children, harms the agricultural sector and perpetuates rural poverty.	FAO supports governments to ensure that child labour issues are better integrated into national agriculture development policies and strategies.
Force-feeding should be banned	The IMA argued against the amendment on the grounds that force-feeding can pose a serious danger to the prisoner's health and violates the ethical rule of doing no harm.	In Washington, Senate leaders continue efforts to force-feed an unpopular Obamacare repeal that will eliminate health coverage for 1.3 million North Carolinians who are now covered.
We should abandon Valentine's day	The Canadian polling firm Insights West surveyed a representative sample of Canadians who are in a relationship and found that 62 percent agreed that Valentine's Day is a waste of time and money.	A recent survey by Virgin Mobile USA found that 59 percent of people said that if they were going to break up with someone, they would do so just before Valentine's Day to save money.

Figure 6: Example of sentences containing similar terms, of which only one is a relevant evidence. Similar terms are in the same color.

L Ein-Dor et al, *Corpus Wide Argument Mining—A Working Solution*, AAAI 2020

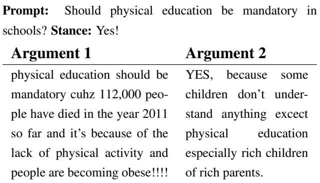
Argument Generation

Civic engagement scenario: arguments from large audiences on debatable topics to generate meaningful narratives. Which arguments to select?

- How to assess argument quality?
- How to summarize **key points** of debate?
  - Deep question with clear connections with NLI
  - e.g., "Women should be able to fight if they are strong enough" and "Women should be able to serve in combat if they choose to" share a large portion of the sentence, but not the main point
  - When are two arguments the same?
  - Crucial for bridging the gap between argument mining and computational argumentation

C Egan et al, *Summarising the points made in online political debates*, ArgMining 2016
H Wachsmuth et al, *Computational Argumentation Quality Assessment in Natural Language*, EACL 2017
A Toledo et al, *Automatic Argument Quality Assessment - New Datasets and Methods*, EMNLP 2019
X Hua & L Wang, *Neural Argument Generation Augmented with Externally Retrieved Evidence*, ACL 2019
R Bar-Haim et al, *From Arguments to Key Points: Towards Automatic Argument Summarization*, ACL 2020

Argument Convincingness



Prompt: Should physical education be mandatory in schools? **Stance:** Yes!

Argument 1	**Argument 2**
physical education should be mandatory cuhz 112,000 people have died in the year 2011 so far and it's because of the lack of physical activity and people are becoming obese!!!!	YES, because some children don't understand anything excect physical education especially rich children of rich parents.

Figure 1: Example of an argument pair.

- What makes an argument persuasive?
- What makes evidence convincing?
  - words related to argumentation (argue, claim), studies, polls, authoritative figures, court orders
  - opinion words (support, opposes, vote), partial change (reduce, amend, part), non-emphasized actions (said, proposed, concern)

I Habernal & I Gurevych, *Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM*, ACL 2016

I Habernal & I Gurevych, *What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation*, EMNLP 2017

I Persing & V Ng, *Why Can't You Convince Me? Modeling Weaknesses in Unpersuasive Arguments*, IJCAI 2017

Gleize et al., *Are you convinced? choosing the more convincing evidence with a siamese network*, ACL 2019

T Chakrabarty et al., *AMPERSAND: Argument Mining for PERSuAsive oNline Discussions*, EMNLP 2019

Argument Reconstruction

Widely recognized as one of the hardest AM tasks

- enthymemes make annotation and automatic analysis challenging
- is it possible for humans to reliably reconstruct?
- implicit premise vs conclusion
- strong dependency on underlying argument model
- the more complex the argument scheme, the more blanks to fill
- Argument Reasoning Comprehension Task

C Stab & I Gurevych, *Parsing Argumentation Structures in Persuasive Essays*, COLI 2017
A Lytos et al., *The evolution of argumentation mining: From models to social media and emerging tools*, Information Processing and Management, 2019
O Razuvayevskaya & S Teufel, *Finding enthymemes in real-world texts: A feasibility study*, Argument and Computation 2017
F Boltuzic & J Šnajder, *Fill the Gap! Analyzing Implicit Premises between Claims from Online Debates*, ArgMining 2016
P Rajendran et al, *Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews*, ArgMining 2016
M Alshomary et al, *Target Inference in Argument Conclusion Generation*, ACL 2020
I Habernal et al, *The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants*, NAACL 2019
T Niven & H-Y Kao, *Probing Neural Network Comprehension of Natural Language Arguments*, ACL 2019

Contemporary AM: Surfing and Scuba Diving

Perspectives

- Improve generalization across corpora and domains
- Address the problem of multilingualism
- Improve scalability of quality labeling
- Properly address structure prediction
- Shift from NLP to NLU possible?
- Move from detection/classification to reasoning in context
- Address fundamental questions: what is the essence of an argument? what makes it persuasive? when are two arguments the same?
- Exploit available AM systems in related tasks

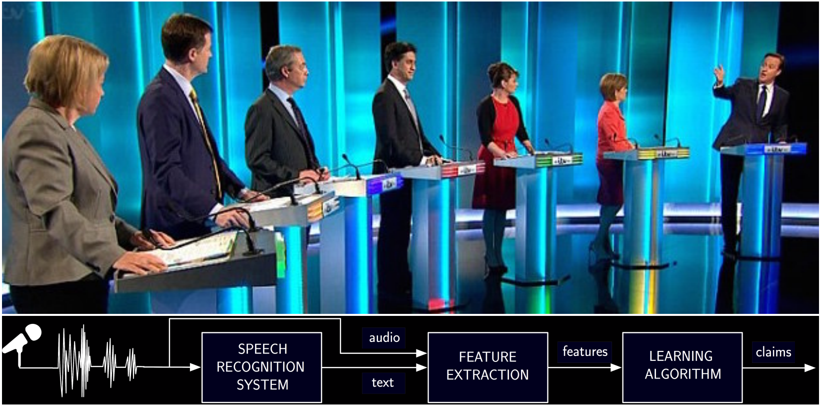
I Gurevych, Latest News in Computational Argumentation: Surfing on the Deep Learning Wave, Scuba Diving in the Abyss of Fundamental Questions, ACL 2017

Related Tasks

Incomplete list of tasks that could benefit from AM

- Study persuasiveness
- Detect the stance of a statement
- Perform fact checking
- Analyze the rhetoric and ethos of a debate
- Retrieve or even generate arguments
- Understanding peer reviews
- Predicting the helpfulness of product reviews
- Improve dialogue systems
- Study citations and scientific argumentation
- Validating argumentation-capable agent-based models
- Speech mining, exploiting both audio and text

Televised Debates



SPEECH RECOGNITION SYSTEM

audio

text

FEATURE EXTRACTION

features

LEARNING ALGORITHM

claims

M Lippi & P Torroni, *Argument Mining from Speech: Detecting Claims in Political Debates*, AAAI 2016

Conclusions

Argumentation mining is a hot topic in AI:

- **very challenging** task
- a lot of **connections** between different sub-areas
- **many potential applications**
- there is still **a lot to be done**
- traction from amazing **advancements in NLP**
- some **working solutions** already available
- great effort is needed to produce **new corpora**
- focus on **general**, less genre-specific AM is an important target

Think big: many problems really at the core of AI:

- understanding **natural language**
- a step **beyond sentiment analysis**
- interaction with **computational** and **natural argumentation**
- learn to **digest information** and reason

From breakthroughs in NLP to breakthrough in argumentation? ☺

Further reading

Selected surveys and references therein:

- A Lytos et al., *The evolution of argumentation mining: From models to social media and emerging tools*, Information Processing and Management, 2019
- J Lawrence & C Reed, *Argument mining: A survey*, Computational Linguistics, 2019
- E Cabrio & S Villata, *Five years of argument mining: A data-driven analysis*, IJCAI, 2018
- M Lippi & P Torroni, *Argumentation mining: State of the art and emerging trends*, ACM Transactions on Internet Technology, 2016
- A Peldszus & M Stede, *From argument diagrams to argumentation mining in texts: A survey*, International Journal of Cognitive Informatics and Natural Intelligence, 2013

Publication venues (mostly open access): IJCAI, AAAI, ECAI, COMMA, ArgMining, ACL, EMNLP, NAACL, EACL, COLING, LREC, journals like Argument & Computation, TACL, COLI, TOIT, AIJ, JAIR