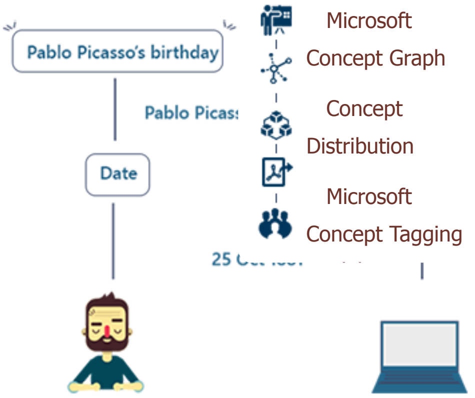
Microsoft Concept Graph Preview

For Short Text Understanding

Microsoft Concept Graph



Microsoft Concept Graph
Concept Distribution
Microsoft Concept Tagging

Our goal is to enable machines to better understand human communication. An important question is, what does the word “understand” mean here? Consider the following example. For human beings, when we see “25 Oct 1881”, we recognize it as a date, although most of us do not know what it is about. However, if we are given a little more context, say the date is embedded in the following piece of short text “Pablo Picasso, 25 Oct 1881, Spain”, most of us would have guessed (correctly) that the date represents Pablo Picasso’s birthday. We are able to do this because we possess certain knowledge, and in this case, “one of the most important dates associated with a person is his birthday.”

As another example, consider a problem in natural language processing. Humans do not find sentences such as “animals other than dogs such as cats” ambiguous, but machine parsing can lead to two possible understandings: “cats are animals” or “cats are dogs.” Common sense tells us that cats cannot be dogs, which renders the second parsing improbal

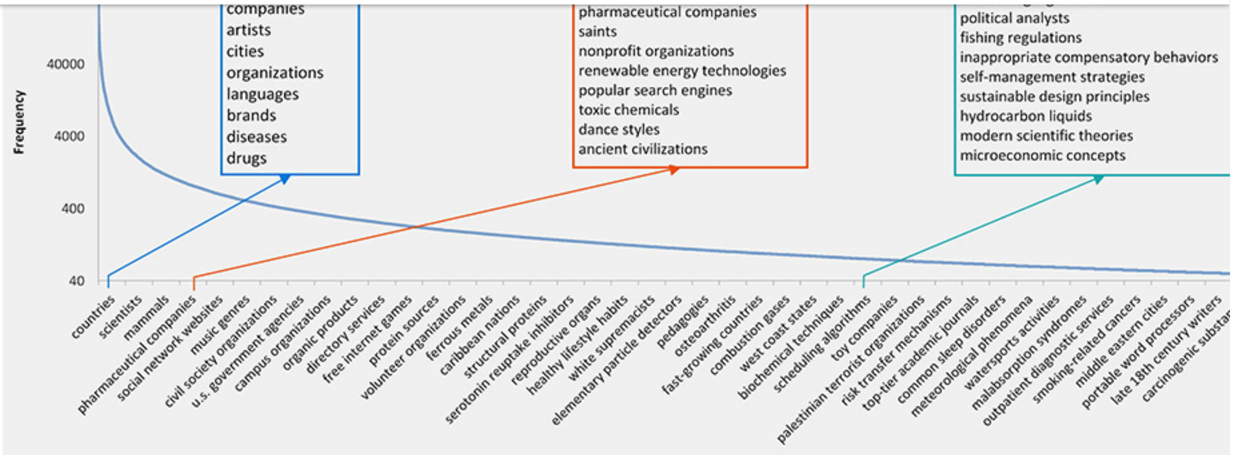
It turns out what we need in order to act like a human in the above two examples is nothing more than knowledge about concept (e.g., persons and animals) and the ability to conceptualize (e.g., cats are animals). This is not a coincidence. Psychologist Gregory Murphy began his highly acclaimed book with the statement “**Concepts are the glue that holds our mental world together**” Nature magazine book review pointed out “Without concepts, there would be no mental world in the first place”. Doubtless to say, having concepts and the ability to conceptualize is one of the defining characteristics of humanity. The question is then: How do w pass human concepts to machines, and how do we enable machines to conceptualize?

In Microsoft Research, we built a research project called Probase, whic big graph of concepts. Knowledge in Probase is harnessed from billion web pages and years' worth of search logs -- these are nothing more t the digitized footprints of human communication. In other words, Prob uses the world as its model. This Microsoft Concept Graph release is bu upon Probase.

Please go to the DOWNLOAD page to get the Microsoft Concept Graph

Concept Distribution

Our mental world contains many concepts about worldly facts, and the Microsoft Concept Graph tries to duplicate them. The core taxonomy of Microsoft Concept Graph alone contains above 5.4 million concepts. The above figure shows their distribution. The Y axis is the number of instances each concept contains(logarithmic scale and on the X axis are the 5.4 million concepts ordered by their size. In contrast, existing knowledge bases have far fewer concept (Freebase contains no more than 2,000 concepts, and Cyc has about 120,000 concepts), which fall short of modeling our mental v As we can see in the above figure, besides popular concepts such as “cities” and “musicians”, which are included by almost every general purpose taxonomy, Microsoft Concept Graph has millions of long tail concepts such as “anti-parkinson treatments”, "celebr



wedding dress designers” and “basic watercolor techniques”, which cannot be found in Freebase or Cyc. Besides concepts, Microsc Concept Graph also has a large data space (each concept contains a set of instances or sub-concepts), a concept is described by a set of attributes), and a large relationship space (e.g., “locatedIn”, "friendOf", " relationships that are not easily named, such as the relationship between apple and Newton.)

In the first release, the Microsoft Concept Graph majorly contains the IsA relation.

Microsoft Concept Graph
Concept Distribution
Microsoft Concept Tagging

Microsoft Concept Tagging Model

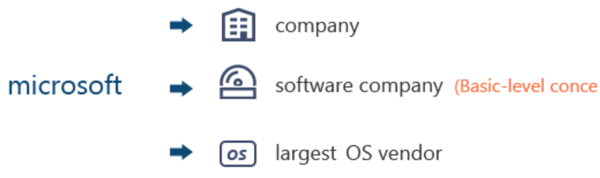
The Microsoft Concept Tagging model (a.k.a. the Conceptualization model) aims to map text format entities into semantic cor categories with some probabilities, which may depend on the context texts of the entities. As an example, “Microsoft” could be automatically mapped to “Software Company” and “Fortune 500 company” etc. with some probabilities. It provides computers the common sense computing capability and make machines "aware" of the mental world of human beings, through which way machi can better understand human communication in text. In detail, conceptualization maps instances or short texts into a large auto learned concept space, which is a vector space, with human-level concept reasoning. It can be treated as both human understand and machine understandable text embedding. Thus it provides us the capability of text concept tagging, short text semantic simila computation etc. for text understanding. It can benefit various text processing applications including search engines, automatic question-answering, online advertising, recommendation systems and artificial intelligence system.

1.Single instance conceptualization (This release)

Single instance conceptualization can return a ranked list of automatically learned concept/category names for any input entity mention/instance. Each concept has a probability to denote the possibility of the input entity belonging to this concept. As a result input entity is represented as a numerical vector, which shows its distribution over the concept vector space.

For human beings, given a single instance, this concept distribution often forms automatically and subconsciously. More importantl those categories at the appropriate level rank higher. Psychologists and linguists call it as **Basic-level Categorization (BLC)**.

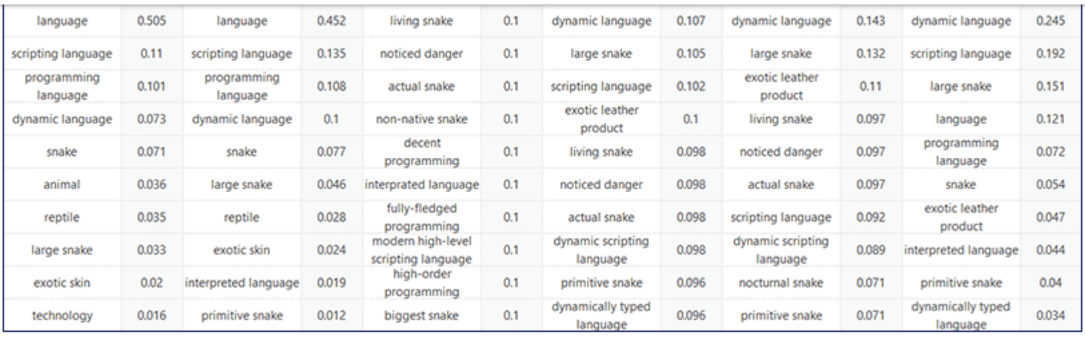
As an example, consider the term ***Microsoft***, which can be categorized into a large number of concepts, ranging from extremely general to extremely specific, such as ***company***, ***software company***, *and* ***largest OS vendor***. If we go through *company*, we may find objects such as McDonald’s and BMW, which have not much similarity to Microsoft. If we go through *largest OS vendor*, we may not be able to find any reasonable object other than Microsoft. On the other hand, if we go through *software company*, we may find Oracle, Adobe, IBM, which are a lot more similar to Microsoft. Thus, software company is a more appropriate basic-level concept for Microsoft, or in oth words, properties associated with *software company* are more readily applied to Microsoft, which is also the reason why through *software company* we can find many objects that are similar to Microsoft.



In this release, we will provide the concept distribution of input text with basic-level conceptualization. Besides, some common measures for conceptualization including MI, PMI, PMIk, and Typicality will be provided simultaneously.

A snapshot of the demo:

Given a single instance “python”, the demo returns concept distributions with different measures (including **BLC** measure):



language	0.505	language	0.452	living snake	0.1	dynamic language	0.107	dynamic language	0.143	dynamic language	0.245
scripting language	0.11	scripting language	0.135	noticed danger	0.1	large snake	0.105	large snake	0.132	scripting language	0.192
programming language	0.101	programming language	0.108	actual snake	0.1	scripting language	0.102	exotic leather product	0.11	large snake	0.151
dynamic language	0.073	dynamic language	0.1	non-native snake	0.1	exotic leather product	0.1	living snake	0.097	language	0.121
snake	0.071	snake	0.077	decent programming	0.1	living snake	0.098	noticed danger	0.097	programming language	0.072
animal	0.036	large snake	0.046	interprated language	0.1	noticed danger	0.098	actual snake	0.097	snake	0.054
reptile	0.035	reptile	0.028	fully-fledged programming	0.1	actual snake	0.098	scripting language	0.092	exotic leather product	0.047
large snake	0.033	exotic skin	0.024	modern high-level scripting language	0.1	dynamic scripting language	0.098	dynamic scripting language	0.089	interpreted language	0.044
exotic skin	0.02	interpreted language	0.019	high-order programming	0.1	primitive snake	0.096	nocturnal snake	0.071	primitive snake	0.04
technology	0.016	primitive snake	0.012	biggest snake	0.1	dynamically typed language	0.096	primitive snake	0.071	dynamically typed language	0.034

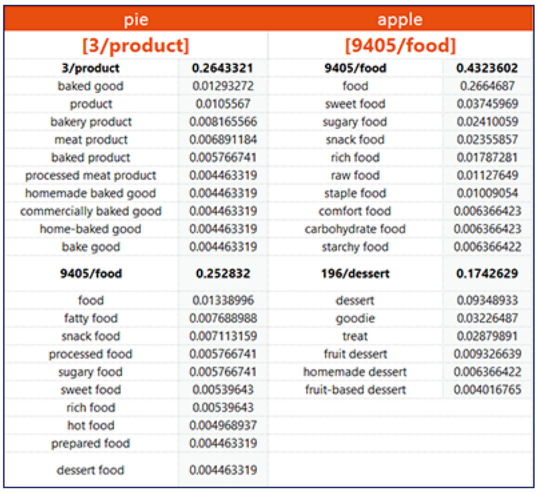
You can simply integrate this single instance conceptualization service into your own applications.

2.Single instance conceptualization with context (v2 release in future)

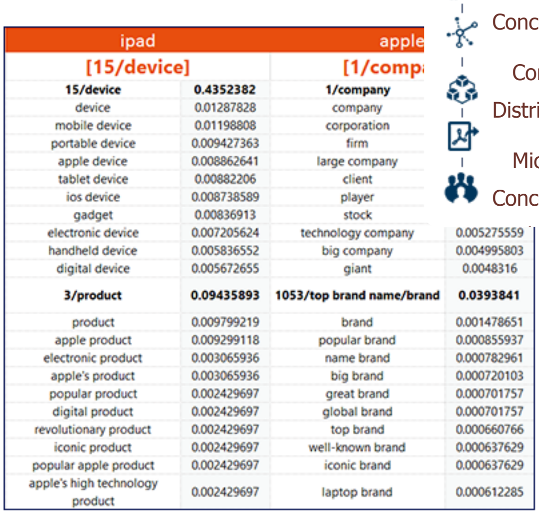
Given “apple” and “pie”, our API maps “apple” to fruit related senses.

Given “apple” and “ipad”, our API maps “apple” to company related seneses.

Microsoft Concept Graph
Concept Distribution
Microsoft Concept Tagging



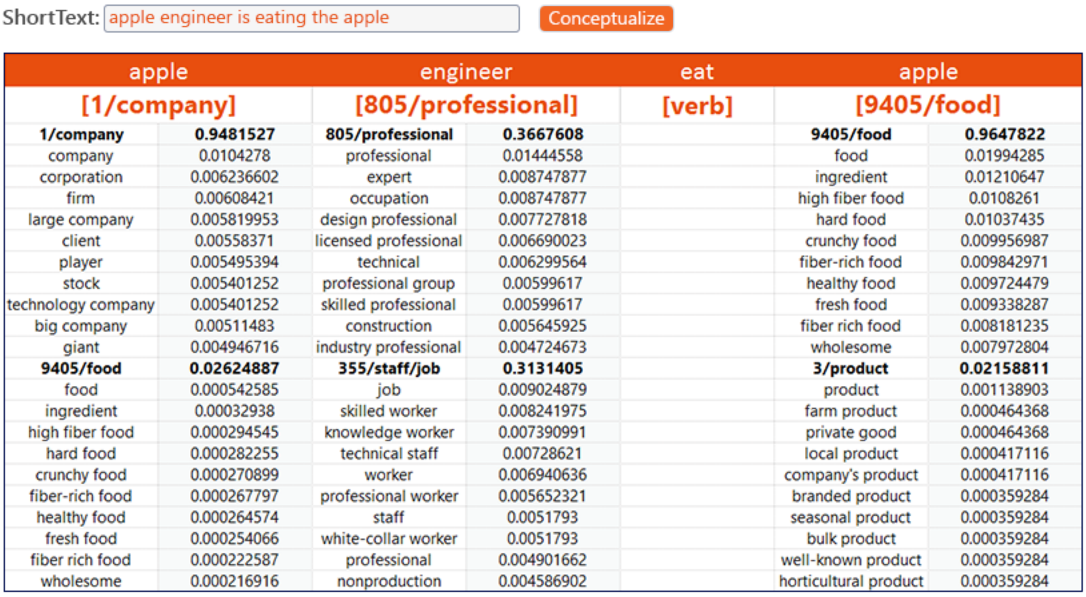
pie		apple	
[3/product]		[9405/food]	
3/product	**0.2643321**	**9405/food**	**0.4323602**
baked good	0.01293272	food	0.2664687
product	0.0105567	sweet food	0.03745969
bakery product	0.008165566	sugary food	0.02410059
meat product	0.006891184	snack food	0.02355857
baked product	0.005766741	rich food	0.01787281
processed meat product	0.004463319	raw food	0.01127649
homemade baked good	0.004463319	staple food	0.01009054
commercially baked good	0.004463319	comfort food	0.006366423
home-baked good	0.004463319	carbohydrate food	0.006366423
bake good	0.004463319	starchy food	0.006366422
9405/food	**0.252832**	**196/dessert**	**0.1742629**
food	0.01338996	dessert	0.09348933
fatty food	0.007688988	goodie	0.03226487
snack food	0.007113159	treat	0.02879891
processed food	0.005766741	fruit dessert	0.009326639
sugary food	0.005766741	homemade dessert	0.006366422
sweet food	0.00539643	fruit-based dessert	0.004016765
rich food	0.00539643		
hot food	0.004968937		
prepared food	0.004463319		
dessert food	0.004463319		



ipad		apple	
[15/device]		[1/comp	
15/device	**0.4352382**	**1/company**	
device	0.01287828	company	
mobile device	0.01198808	corporation	
portable device	0.009427363	firm	
apple device	0.008862641	large company	
tablet device	0.00882206	client	
ios device	0.008738589	player	
gadget	0.00836913	stock	
electronic device	0.007205624	technology company	0.005275559
handheld device	0.005836552	big company	0.004995803
digital device	0.005672655	giant	0.0048316
3/product	**0.09435893**	**1053/top brand name/brand**	**0.0393841**
product	0.009799219	brand	0.001478651
apple product	0.009299118	popular brand	0.000855937
electronic product	0.003065936	name brand	0.000782961
apple's product	0.003065936	big brand	0.000720103
popular product	0.002429697	great brand	0.000701757
digital product	0.002429697	global brand	0.000701757
revolutionary product	0.002429697	top brand	0.000660766
iconic product	0.002429697	well-known brand	0.000637629
popular apple product	0.002429697	iconic brand	0.000637629
apple's high technology product	0.002429697	laptop brand	0.000612285

3.Short text conceptualization (v3 release in future)

Given a short text “the engineer is eating the apple”, will do the segmentation, concept mapping, and sense disambiguation.



ShortText: apple engineer is eating the apple

apple		engineer		eat	apple	
[1/company]		[805/professional]		[verb]	[9405/food]	
1/company	**0.9481527**	**805/professional**	**0.3667608**		**9405/food**	**0.9647822**
company	0.0104278	professional	0.01444558		food	0.01994285
corporation	0.006236602	expert	0.008747877		ingredient	0.01210647
firm	0.00608421	occupation	0.008747877		high fiber food	0.0108261
large company	0.005819953	design professional	0.007727818		hard food	0.01037435
client	0.00558371	licensed professional	0.006690023		crunchy food	0.009956987
player	0.005495394	technical	0.006299564		fiber-rich food	0.009842971
stock	0.005401252	professional group	0.00599617		healthy food	0.009724479
technology company	0.005401252	skilled professional	0.00599617		fresh food	0.009338287
big company	0.00511483	construction	0.005645925		fiber rich food	0.008181235
giant	0.004946716	industry professional	0.004724673		wholesome	0.007972804
9405/food	**0.02624887**	**355/staff/job**	**0.3131405**		**3/product**	**0.02158811**
food	0.000542585	job	0.009024879		product	0.001138903
ingredient	0.00032938	skilled worker	0.008241975		farm product	0.000464368
high fiber food	0.000294545	knowledge worker	0.007390991		private good	0.000464368
hard food	0.000282255	technical staff	0.00728621		local product	0.000417116
crunchy food	0.000270899	worker	0.006940636		company's product	0.000417116
fiber-rich food	0.000267797	professional worker	0.005652321		branded product	0.000359284
healthy food	0.000264574	staff	0.0051793		seasonal product	0.000359284
fresh food	0.000254066	white-collar worker	0.0051793		bulk product	0.000359284
fiber rich food	0.000222587	professional	0.004901662		well-known product	0.000359284
wholesome	0.000216916	nonproduction	0.004586902		horticultural product	0.000359284

References

Please cite following papers if you use our data:

October 2015.

2. *Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Zhu, Probase: A Probabilistic Taxonomy for Text Understanding, in ACM International Conference on Management of Data (SIGMOD), May 2012.*

Please cite following papers if you use our conceptualization service:

1. *Zhongyuan Wang and Haixun Wang, Understanding Short Texts, in the Association for Computational Linguistics (ACL) (Tuto August 2016.*
2. *Zhongyuan Wang, Haixun Wang, Ji-Rong Wen, and Yanghua Xiao, An Inference Approach to Basic Level of Categorization, i ACM International Conference on Information and Knowledge Management (CIKM), ACM –Association for Computing Machi October 2015.*
3. *Zhongyuan Wang, Kejun Zhao, Haixun Wang, Xiaofeng Meng, and Ji-Rong Wen, Query Understanding through Knowledge-B Conceptualization, in IJCAI, July 2015.*
4. *Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou, Short Text Understanding Through Lexical-Sema Analysis, in International Conference on Data Engineering (ICDE), April 2015. (**Best Paper Award**)*
5. *Zhongyuan Wang, Haixun Wang, and Zhirui Hu, Head, Modifier, and Constraint Detection in Short Texts, in International Conference on Data Engineering (ICDE), 2014.*
6. *Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen, Short Text Conceptualization using a Probabilistic Knowledgebase, in IJCAI, 2011.*

Microsoft Concept Graph

Concept Distribution

Microsoft Concept Tagging

Contacts

Team Members:

Yaobo Liang Lei Ji

Group

Data Mining and Enterprise Intelligence Group, MSRA

Acknowledgments

We would like to acknowledge **Haixun Wang, Zhongyuan Wang, Dawei Zhang, Jun Yan, Yangqiu Song, Hongsong Li, an many interns** for their contributions to the Microsoft Concept Graph and the Microsoft Concept Tagging model. Especially for **Hai Wang**, he initiated and led this project when he was at Microsoft Research. We highly appreciate his tremendous contributions an insightful vision which make this project succeed finally.