A Cross-Lingual Dictionary for English Wikipedia Concepts

Valentin I. Spitkovsky, Angel X. Chang

Google Research, Google Inc., Mountain View, CA, 94043
Computer Science Department, Stanford University, Stanford, CA, 94305
{valentin, angelx}@{google.com, cs.stanford.edu}

Abstract

We present a resource for automatically associating strings of text with English Wikipedia concepts. Our machinery is bi-directional, in the sense that it uses the same fundamental probabilistic methods to map strings to empirical distributions over Wikipedia articles as it does to map article URLs to distributions over short, language-independent strings of natural language text. For maximal interoperability, we release our resource as a set of flat line-based text files, lexicographically sorted and encoded with UTF-8. These files capture joint probability distributions underlying concepts (we use the terms **article**, **concept** and Wikipedia **URL** interchangeably) and associated snippets of text, as well as other features that can come in handy when working with Wikipedia articles and related information.

Keywords: cross-language information retrieval (CLIR), entity linking (EL), Wikipedia.

1. Introduction

Wikipedia's increasingly broad coverage of important concepts brings with it a valuable high-level structure that organizes this accumulated collection of world knowledge. To help make such information even more "universally accessible and useful," we provide a mechanism for mapping between Wikipedia articles and a lower-level representation: free-form natural language strings, in many languages. Our resource's quality was vetted in *entity linking* (EL) competitions, but it may also be useful in other *information retrieval* (IR) and *natural language processing* (NLP) tasks.

2. The Dictionary

The resource that we constructed closely resembles a dictionary, with canonical English Wikipedia URLs on the one side, and relatively short natural language strings on the other. These strings come from several disparate sources, primarily: (i) English Wikipedia titles; (ii) anchor texts from English inter-Wikipedia links; (iii) anchor texts into the English Wikipedia from non-Wikipedia web-pages; and (iv) anchor texts from non-Wikipedia pages into non-English Wikipedia pages, for topics that have corresponding English Wikipedia articles. Unlike entries in traditional dictionaries, however, the strengths of associations between related pairs in our mappings can be quantified, using basic statistics. We have sorted our data using one particularly simple scoring function (a conditional probability), but we include all raw counts so that users of our data could experiment with metrics that are relevant to their specific tasks.[1]

3. High-Level Methodology

Our scoring functions $\mathbb{S}$ are essentially conditional probabilities: they are ratios of the number of hyper-links into a Wikipedia URL having anchor text s and either (i) the total number of anchors with text s, $\mathbb{S}(\text{URL} \mid s)$, for going from strings to concepts; or (ii) the count of all links pointing to an article, $\mathbb{S}(s \mid \text{URL})$, for going from concepts to strings.

[1]Web counts are from a subset of a 2011 Google crawl.

Zero scores are added in, explicitly, for article titles and other relevant strings that have not been seen in a web-link.

Further details about the components of these scoring functions are outlined in our earliest system description paper (Agirre et al., 2009, §2.2). Many other low-level implementation details are in the rest of its section about the dictionary (Agirre et al., 2009, §2) and in the latest, cross-lingual system description (Spitkovsky and Chang, 2011).

4. From Strings to Concepts

Let us first discuss using the dictionary as a mapping from strings s to canonical URLs of English Wikipedia concepts. Table 1 shows the scores of all entries that match the string *Hank Williams* — a typical *entity linking* (EL) task (McNamee and Dang, 2009; Ji et al., 2010) query — exactly. We see in these results two salient facts: (i) the dictionary exposes the ambiguity inherent in the string *Hank Williams* by distributing probability mass over several concepts, most of which have some connection to one or another Hank

$\mathbb{S}(\text{URL} \mid s)$	*Canonical (English)* URL
0.990125	Hank_Williams
0.00661553	Your_Cheatin'_Heart
0.00162991	Hank_Williams,_Jr.
0.000479386	I
0.000287632	Stars_&_Hank_Forever: _The_American_Composers_Series
0.000191755	I'm_So_Lonesome_I_Could_Cry
0.000191755	I_Saw_the_Light_(Hank_Williams_song)
0.0000958773	Drifting_Cowboys
0.0000958773	Half_as_Much
0.0000958773	Hank_Williams_(Clickradio_CEO)
0.0000958773	Hank_Williams_(basketball)
0.0000958773	Lovesick_Blues
0	Hank_Williams_(disambiguation)
0	Hank_Williams_First_Nation
0	Hank_Williams_III
1.0	

Table 1: All fifteen dictionary entries matching the string $s =$ *Hank Williams* exactly (the raw counts are not shown).

Williams; and (ii) the dictionary effectively disambiguates the string, by concentrating most of its probability mass on a single entry. These observations are in line with similar insights from the *word sense disambiguation* (WSD) literature, where the "most frequent sense" (MFS) serves as a surprisingly strong baseline (Agirre and Edmonds, 2006).[2]

5. From Concepts to Strings

We now consider running the dictionary in reverse. Since anchor texts that link to the same Wikipedia article are coreferent, they may be of use in coreference resolution and, by extension (Recasens and Vila, 2010), paraphrasing. For our next example, we purposely chose a concept that is not a named entity: `Soft_drink`. Because the space of strings is quite large, we restricted the output of the dictionary, excluding strings that originate only from non-Wikipedia pages and strings landing only on non-English articles (see Table 2), by filtering on the appropriate raw counts (which are included with the dictionary). We see in this table a noisy but potentially useful data source for mining synonyms (for clarity, we aggregated on punctuation, capitalization and pluralization variants). Had we included all dictionary entries, there would have been even more noise, but also translations and other varieties of natural language text referring to similar objects in the world.

6. An Objective Evaluation

The *entity linking* (EL) task — as defined in Knowledge-Base Population (KBP) tracks at the Text Analysis Conferences (TACs) — is a challenge to disambiguate string mentions in documents. Ambiguity is to be resolved by associating specific mentions in text to articles in a knowledge base (KB, derived from a subset of Wikipedia). We evaluated the dictionary by participating in all (English) TAC-KBP entity linking challenges (Agirre et al., 2009; Chang et al., 2010; Chang et al., 2011), as well as in the most recent cross-lingual bake-off (Spitkovsky and Chang, 2011).

English-only versions of the dictionary have consistently done well — scoring above the median entry — in all three monolingual competitions.[3] The reader may find this surprising, as did we, considering that the dictionary involves no machine learning (i.e., we did not tune any weights) and is entirely context-free (i.e., uses only the query to perform a look-up, ignoring surrounding text) — i.e., it is a baseline.

In the cross-lingual bake-off, perhaps not surprisingly, the English-only dictionary scored below the median; however, the full cross-lingual dictionary once again outperformed more than half of the systems, despite its lack of supervision, a complete disregard for context, and absolutely no language-specific adaptations (in that case, for Chinese).

In-depth quantitative and qualitative analyses describing the latest challenge are available in a report (Ji et al., 2011) furnished by the conference's organizers.

[2]First-sense heuristics are also (transitively) used in work outside WSD, such as ontology merging — e.g., in YAGO (Suchanek et al., 2008), combining Wikipedia with WordNet (Miller, 1995).

[3]Using a simple disambiguation strategy on top of the dictionary, our submission to the 2010 contest scored higher than all other systems not accessing recently updated Wikipedia pages.

$\mathbb{S}(s \mid \text{URL})$	*String* s	*(and Variants)*
0.2862316	soft drink	*(and* soft-drinks*)*
0.0544652	soda	*(and* sodas*)*
0.00858187	soda pop	
0.00572124	fizzy drinks	
0.003200497	carbonated beverages	*(and* beverage*)*
0.002180871	non-alcoholic	
0.00141615	soft	
0.001359502	pop	
0.001132923	carbonated soft drink	*(and* drinks*)*
0.000736398	aerated water	
0.000708075	non-alcoholic drinks	*(and* drink*)*
0.000396522	soft drink controversy	
0.000311553	citrus-flavored soda	
0.00028323	carbonated	
0.000226584	soft drink topics	
0.000226584	carbonated drinks	
0.000198261	soda water	
0.000169938	grape soda	
0.000113292	juice drink	
0.000113292	sugar-sweetened drinks	
0.000084969	beverage	
0.000084969	lemonades	*(and* lemonade*)*
0.000056646	flavored soft drink	
0.000056646	pop can	
0.000056646	obesity and selling soda to children	
0.000028323	cold beverages	
0.000028323	fizzy	
0.000028323	other soft drinks	
0.000028323	beverage manufacturer	
0.000028323	health effects	
0.000028323	minerals	
0.000028323	onion soda	
0.000028323	soda drink	
0.000028323	soft beverage	
0.000028323	tonics	
0.3683967		

Table 2: Dictionary scores for anchor text strings that refer to the URL `Soft_drink` within the English Wikipedia, after normalizing out capitalization, pluralization and punctuation; note that nearly two thirds (63.2%) of web links have anchor text that is unique to non-English-Wikipedia pages.

$\mathbb{S}(\text{URL} \mid s)$	URL	*(and Associated Scores)*
0.966102	`Galago`	D W:110/111 W08 W09 WDB w:2/5 w':2/2
0.0169492	`bushbaby`	w:2/5
0.00847458	`Lesser_bushbaby`	W:1/111 W08 W09 WDB
0.00847458	`bushbabies`	c t w:1/5

Table 3: All dictionary entries for string $s =$ *bushbabies*. The top result is linked from a disambiguation page (D) and absorbs 110 of all 111 web-links (W) into English Wikipedia with this anchor text; it also takes two of the five inter-English-Wikipedia links (w), based on information in our Wikipedia dumps from 2008, 2009 and DBpedia (W08, W09 and WDB) — two of two, based on a more recent Google crawl (w'). Its score is $114/118 \approx 96.6\%$. The last result is in a cluster with Wikipedia pages (itself) having s as both a title (t) and consequently a clarification (c). Absence of counts from non-English Wikipedia pages (Wx) confirms that results are English-only (boolean x not set).

7. Some Examples and Low-Level Details

The dictionary will be distributed as a static resource,[4] serialized over seven files. Its key objects are English Wikipedia URLs, non-empty strings s and their so-called "LNRM" (Agirre et al., 2009, §2.3) forms $l(s)$, which are canonical representations that ignore white-space, tolerate case differences, various font and diacritic variations, etc. In addition to these three types of objects, the dictionary contains mapping scores, raw counts, and many other features suitable for use with machine learning algorithms.

- `dictionary`: maps strings s to canonical URLs — see Table 3 for a detailed example;
- `inv.dict`: maps canonical URLs back to strings s — see Tables 4–7 for detailed examples;
- `cross.map`: maps non-English to canonical URLs — e.g., `de.wikipedia.org/wiki/Riesengalagos` to `Greater_galago`;
- `redir.map`: maps free-style titles to canonical URLs — e.g., *Bush Baby* and *Bushbabies* to `Greater_galago`;
- `lnrm.forw`: maps strings s to canonical $l(s)$ — e.g., *Bushbaby (lesser)* to `bushbabylesser`;
- `lnrm.back`: maps strings $l(s)$ back to s — e.g., `bushbabylesser` to *Bushbaby (lesser)*, etc.
- `lnrm.dict`: maps aggregate $l(s)$ to canonical URLs.

An eighth file, `redir.log`, contains a trace of all proposed cluster merges, which resulted from executing the *union-find* (UF) algorithm over dozens of relaxations of Wikipedia redirects graphs, before finally yielding `redir.map`.

8. Related Work

Our resource is not the first tool for mapping between text strings and Wikipedia concepts. For example, Milne and Witten (2008) trained a system to inject hyper-links into Wikipedia-like text. And still earlier, Gabrilovich and Markovitch (2007) exploited Wikipedia concepts as a low-dimensional representation for embedding natural language, via *explicit components analysis* (ESA) of "bag of words" (BOW) models. Previous approaches heavily relied on the actual text in Wikipedia articles, which vary wildly, both in the quantity and quality of their content.

An early study (Giles, 2005) that compared the quality of scientific articles in Wikipedia with those of *Encyclopædia Britannica* found that the difference was "not particularly great," stirring a fair bit of controversy.[5] But even academics who argue against classifying Wikipedia with traditional encyclopedias emphasize its increasing use as a source of shared information (Magnus, 2006). Our systems leverage precisely this wide-spread use — and not the intrinsic quality or size — of Wikipedia's articles by associating anchor texts (collected by crawling a reasonably large approximation of the entire web) with Wikipedia's broad-coverage span of important concepts and relevant topics.

The dictionary is most similar to the work of Koningstein et al. (2003a; 2003b; 2004), which connected search engine advertising keywords with vertical sales categories. The main differences lie in using (i) Wikipedia concepts in place of the Open Directory Project (ODP) categories; and (ii) publicly-available anchor text of links into Wikipedia instead of proprietary queries of click-throughs to ODP.[6]

9. Summary of Contributions

The dictionary is a large-scale resource which would be difficult to reconstruct in a university setting, without access to a comprehensive web-crawl. It offers a strong baseline for entity linking, but primarily through sheer engineering effort. In releasing the data, we hope to foster new advances, by allowing research focus to shift firmly towards context-sensitive and machine learning methods that would build on top of its large volume of information (Halevy et al., 2009).[7] Along with the core dictionary, we release several other useful mappings, including: (i) from non-English Wikipedia URLs to the corresponding English analogs; and (ii) from free-style English Wikipedia titles to the canonical URLs, including active redirects by Wikipedia's servers.

Although we did not carefully evaluate the dictionary for natural language processing tasks other than entity linking, we suspect that it could be of immediate use in many other settings as well. These include some areas that we already mentioned (e.g., paraphrasing and coreference resolution, machine translation and synonym mining), and hopefully many others (e.g., natural language generation). By releasing the dictionary resource, we hope to fuel numerous creative applications that will have been difficult to predict.

10. Acknowledgments

This work was carried out in the summer of 2011, while both authors were employed at Google Inc., over the course of the second author's internship. We would like to thank our advisors, Dan Jurafsky and Chris Manning, at Stanford University, for their continued help and support. We are also grateful to the other members of the original Stanford-UBC TAC-KBP entity linking team — Eneko Agirre and Eric Yeh: our initial (monolingual) dictionary for mapping strings to Wikipedia articles was conceived and constructed during a collaboration with them, in the summer of 2009.

We thank Nate Chambers, Dan Jurafsky, Marie-Catherine de Marneffe and Marta Recasens — of the Stanford NLP Group — and the anonymous reviewer(s) for their help with draft versions of this paper. Last but not least, we are grateful to many Googlers — Thorsten Brants, Johnny Chen, Eisar Lipkovitz, Peter Norvig, Marius Paşca and Agnieszka Purves — for guiding us through the internal approval processes that were necessary to properly release this resource.

[4] `nlp.stanford.edu/pubs/crosswikis-data.tar.bz2`

[5] See *Britannica*'s response and *Nature*'s reply, "Britannica attacks... and we respond," at `corporate.britannica.com/britannica_nature_response.pdf` and `www.nature.com/nature/britannica`, respectively.

[6] `www.dmoz.org`

[7] The dictionary consists of 297,073,139 associations, mapping 175,100,788 unique strings to related English Wikipedia articles.

$\mathbb{S}(s \mid \text{URL})$	*String* s	W (of 8,594)	Wx (of 6,207)	w (of 73)	w' (of 140)
0.24244	ceviche	2,826	724	35	55
0.164113	Ceviche	1,803	564	28	69
0.0644732	http://en.wikipedia.org/wiki/Ceviche	968			
0.0366991	cebiche	36	514		1
0.0326362	Cebiche	132	358		
0.0225123	Ceviche - Wikipedia, the free encyclopedia	338			
0.0212468	ceviches	195	122		2
0.0189823	Cebiche - Wikipedia, la enciclopedia libre		285		
0.0169841	http://de.wikipedia.org/wiki/Ceviche		255		
0.012455	Ceviches de Camaron	187			
0.012122	Wikipedia	119	63		
0.0103903	Wikipedia: Ceviche	156			
0.00972426	http://es.wikipedia.org/wiki/Ceviche		146		
0.00706008	en.wikipedia.org/wiki/Ceviche	106			
0.00679366	http://es.wikipedia.org/wiki/Cebiche		102		
0.00672705	[1]	60	41		
0.00619422	seviche	35	58		
0.00506194	comida peruana		76		
0.00506194	here	38	38		
0.00506194	“ceviche”		76		
0.00492873	Kinilaw	38	32	2	2
0.00472892	[4]	15	56		
0.00426269	Wikipedia.org		64		
0.00419608	(External) ceviche	63			
0.00419608	cebiches	1	62		
0.00399627	sebiche		60		
0.00386306	[3]	22	36		
0.00346343	ceviched	52			
0.00339683	cebichería		51		
0.00333023	セビチェ		50		
0.00319702	Cerviche	42	6		
0.00319702	セビーチェ		48		
0.00313041	Turn to Wikipedia *(in Hebrew)*		47		
0.0029972	с е в и ч е		45		
0.00279739	C - Ceviche in Peru	42			
0.00273078	Ceviche del Perú.jpg		41		
0.00273078	Kilawin	32	7	1	1
0.00266418	セビチェ - Wikipedia		40		
0.00259758	kinilaw	32	2	2	3
0.00253097	Seviche	16	22		
0.00253097	[6]	24	14		
0.00246437	[5]	17	20		
0.00239776	Deutsch		36		
0.00239776	Source: Wikipedia	36			
0.00239776	Svenska		36		
0.00233116	CEVICHE	6	29		
0.00233116	[2]	3	32		
0.00233116	日本語		35		
0.00219795	Hebrew *(in Hebrew)*		33		
0.00213134	Français		32		
0.00213134	http://pl.wikipedia.org/wiki/Ceviche		32		
0.00213134	kilawin	26		2	4
0.00206474	Español		31		
0.00206474	Tagalog		31		
0.00199814	Ceviche de pescado		30		
0.00199814	Peruvian ceviche	18	11	1	
0.8115749	⋮	7,484	4,493	71	137

Table 4: The 56 highest-scoring strings s for Wikipedia URL `Ceviche` — unfiltered and, admittedly, quite noisy: there are many URL strings, mentions of Wikipedia, citation references (e.g., *[1], [2]*, and so on), side comments (e.g., *(External)*), names of languages, the notorious “*here*” link, etc. Nevertheless, the title string *ceviche* is at the top, with alternate spellings (e.g., *cebiche* and *seviche*) and translations (e.g., *kinilaw*) not far behind. Hit counts from the Wikipedia-external web into the English Wikipedia page (W), its non-English equivalents (Wx) and inter-English-Wikipedia links (w, from older English Wikipedia dumps, and w', from a recent Google web-crawl) could be used to effectively filter out some noise.

$\mathbb{S}(s \mid \text{URL})$	*String* s	W	Wx	w	w’
0.00159851	cheviche	3	21		
0.0014653	セビッチェ		22		
0.00139869	El seviche o ceviche		21		
0.00126549	El cebiche		19		
0.00126549	ceviçhe	8	11		
0.00119888	shrimp ceviche	18			
0.00106567	Ceviche (eine Art Fischsalat)		16		
0.00106567	cebiche peruano		16		
0.00106567	cerviche	16			
0.000932463	“Ceviche”	14			
0.000932463	Cebiche peruano		14		
0.000865859	El Ceviche		13		
0.000865859	El ceviche		13		
0.000799254	Ceviche blanco		12		
0.000799254	Juan José Vega		12		
0.00073265	Ceviche:	7	4		
0.00073265	South American ceviche	11			
0.00073265	С е в и ч е	1	10		
0.000666045	Peru...Masters of Ceviche	10			
0.000666045	cevichito		10		
0.000666045	puts their own twist	10			
0.000666045	tiradito		10		
0.000599441	Chinguirito		9		
0.000599441	cevichazo		9		
0.000599441	the right kind	9			
0.000532836	Sebiche		8		
0.000532836	mestizaje y aporte de las diversas culturas		8		
0.000532836	trout ceviche	8			
0.000466232	cevice	7			
0.000466232	el ceviche		7		
0.000466232	le ceviche		7		
0.000466232	leckere Ceviche		7		
0.000399627	Ceviche o cebiche es el nombre de diversos		6		
0.000399627	ceviche peruano		6		
0.000399627	unique variation	6			
0.000333023	“Kinilaw”	5			
0.000333023	“ceviche”		5		
0.000333023	“ceviches”		5		
0.000333023	Cevichen		5		
0.000333023	Spécialité d’Amérique Latine		5		
0.000333023	e che sarebbe ’sto ceviche?		5		
0.000333023	food	5			
0.000333023	kilawing	4			1
0.000333023	o ceviche		5		
0.000333023	“cevichele”	5			
0.000266418	Cebiches		4		
0.000266418	Ceviche Tostada	4			
0.000266418	Ceviche de camarones		4		
0.000266418	Ceviche!	4			
0.000266418	Ceviche, cebiche, seviche o sebiche		4		
0.000266418	El ceviche es peruano		4		
0.000266418	The geeky chemist in me loves “cooking” proteins	4			
0.000266418	You know ceviche	4			
0.000266418	ahi tuna ceviche	4			
0.000266418	ceviche (peruano)		4		
0.000266418	ceviche de pesca	4			
0.000266418	chevichen		4		
0.000266418	civiche	4			
0.000266418	el cebiche		4		
0.000266418	el cebiche o ceviche		4		

⋮

Table 5: A non-random sample of 60 from the next 192 strings (offsets 57 through 248) associated with `Ceviche`.

$\mathbb{S}(s \mid \text{URL})$	*String* s	W	Wx	w	w’
0.000266418	seviches	4			
0.000266418	ϛεβιτϛε	4			
0.000266418	セビッチェ屋		4		
0.000266418	海鮮料理セビッチェ		4		
0.000199814	A PRUEBA DE CEVICHE.		3		
0.000199814	Ceviche de Mariscos		3		
0.000199814	Cevicheria	1	2		
0.000199814	El Día Nacional del Cebiche		3		
0.000199814	It forms a kind of ceviche.	3			
0.000199814	cebiche o ceviche		3		
0.000199814	cebiche ría		3		
0.000199814	cebicheria	1		1	1
0.000199814	ceviche mixo	3			
0.000199814	ceviche style	3			
0.000199814	ceviche!		3		
0.000199814	cevicheria	3			
0.000199814	cevicheriak	3			
0.000199814	chevice	3			
0.000199814	citrus-marinated seafood	3			
0.000199814	es sobre todo de los peruanos		3		
0.000199814	peixe cru com limão e cebola		3		
0.000199814	seafood	3			
0.000199814	メキシコやペルーで食される海産物マリネ「セビーチェ」風		3		
0.000133209	“El Ceviche”	2			
0.000133209	Cebicherias			1	1
0.000133209	Ceviche (selbst noch nicht probiert)		2		
0.000133209	Ceviche de Corvina	2			
0.000133209	Ceviche de Mahi Mahi con platano frito		2		
0.000133209	Ceviche de Pescado		2		
0.000133209	Ceviche de camarón ecuatoriano		2		
0.000133209	Ceviche mixto	2			
0.000133209	Ceviche(セビーチェ）		2		
0.000133209	Ceviches de pescado , pulpo, calamar, langosta y cangrejo		2		
0.000133209	Cevicheセビチェ		2		
0.000133209	Cheviche	2			
0.000133209	Civeche	2			
0.000133209	Civiche	2			
0.000133209	Le Ceviche		2		
0.000133209	Mmmmmmmm.....	2			
0.000133209	Peruvian ceviché	2			
0.000133209	What is the origin of Ceviche?	2			
0.000133209	cerveche	2			
0.000133209	cevi	2			
0.000133209	ceviche de camaron	2			
0.000133209	ceviche de pescado	2			
0.000133209	ceviche de pulpo	2			
0.000133209	ceviche till forratt.		2		
0.000133209	ceviche/cebiche	2			
0.000133209	cevicheä	2			
0.000133209	conchas negras		2		
0.000133209	cooked	2			
0.000133209	exactly what it is	2			
0.000133209	marinated seafood salad	2			
0.000133209	tuna ceviche	2			
0.000133209	un plato de comida		2		
0.000133209	whatever that is	2			
0.000133209	“Cerviche”	2			
0.000133209	『セビチェ』の解説		2		
0.000133209	いろんな具材		2		
0.000133209	セビチェ *(narrow script)*		2		

⋮

Table 6: A non-random sample of 60 from the next 204 strings (offsets 249 through 452) associated with `Ceviche`.

$\mathbb{S}(s \mid \text{URL})$	*String* s	W	Wx	w	w’
0.0000666045	Caviche according to Wikipedia	1			
0.0000666045	Cebiche - Wikipedia		1		
0.0000666045	Ceviche - Authentic Mexican Food Fish Recipe	1			
0.0000666045	Ceviche / Wiki		1		
0.0000666045	Ceviche bei der wikipedia		1		
0.0000666045	Ceviche por país		1		
0.0000666045	Ceviche; it is used under the	1			
0.0000666045	Ceviche?	1			
0.0000666045	Diferentes versiones del cebiche forman parte de la		1		
0.0000666045	En México		1		
0.0000666045	Fish, lemon, onion, chilli pepper. Ceviche[3] (also	1			
0.0000666045	Impacto socio-cultural		1		
0.0000666045	Kinilaw; it is used under the	1			
0.0000666045	La historia del ceviche		1		
0.0000666045	Los Calamarcitos - Ceviche, Comida tipica arequipeña, Mariscos		1		
0.0000666045	On débat de l’étymologie de ceviche		1		
0.0000666045	Peru - Ceviche	1			
0.0000666045	Preparation	1			
0.0000666045	Recette:		1		
0.0000666045	Saviche		1		
0.0000666045	Shrimp Ceviche Recipe	1			
0.0000666045	This dish	1			
0.0000666045	Today ceviche is a popular international dish prepared	1			
0.0000666045	Try this, will blown your tongue away!	1			
0.0000666045	Variations	1			
0.0000666045	Walleye Ceviche	1			
0.0000666045	Wikipedia (Cebiche)		1		
0.0000666045	Wikipedia (Ceviche)	1			
0.0000666045	Wikipedia Entry on Ceviche	1			
0.0000666045	a different food term that can kill you	1			
0.0000666045	airport ceviche	1			
0.0000666045	cebiche exists in		1		
0.0000666045	cebiche)		1		
0.0000666045	cebiche,		1		
0.0000666045	ceviche (the national dish)	1			
0.0000666045	ceviche bar	1			
0.0000666045	ceviche peruano.		1		
0.0000666045	ceviche salsa dip.	1			
0.0000666045	ceviche that she ordered there. After quizzing her	1			
0.0000666045	ceviche tostada	1			
0.0000666045	ceviche y		1		
0.0000666045	ceviche)		1		
0.0000666045	cevichera		1		
0.0000666045	cevishe.	1			
0.0000666045	civiche is okay	1			
0.0000666045	dinner	1			
0.0000666045	dish	1			
0.0000666045	eviche		1		
0.0000666045	o cevich	1			
0.0000666045	raw, marinated in sour lime juice, with onions	1			
0.0000666045	rå fisk marinert i lime, Cebiche		1		
0.0000666045	seviché	1			
0.0000666045	- Kinilaw :	1			
0.0000666045	About Ceviche	1			
0.0000666045	CERVICHE	1			
0.0000666045	CEVICHE DE MARISCO Videos - Pakistan Tube - Watch Free	1			
0.0000666045	Цевишхе	1			
0.0000666045	『セビーチェ』		1		
0.0000666045	セビチェ-wikipedia *(narrow script)*		1		
0	saviche				

Table 7: A non-random sample of 60 from 246 hapax legomena and the last of the zero-scorers associated with `Ceviche`.

Figure 1: The first author dedicates his contribution to Amber, who (to the best of our knowledge) never got to try ceviche.

11. References

E. Agirre and P. Edmonds, editors. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer.

E. Agirre, A. X. Chang, D. S. Jurafsky, C. D. Manning, V. I. Spitkovsky, and E. Yeh. 2009. Stanford-UBC at TAC-KBP. In *TAC*.

A. X. Chang, V. I. Spitkovsky, E. Yeh, E. Agirre, and C. D. Manning. 2010. Stanford-UBC entity linking at TAC-KBP. In *TAC*.

A. X. Chang, V. I. Spitkovsky, E. Agirre, and C. D. Manning. 2011. Stanford-UBC entity linking at TAC-KBP, again. In *TAC*.

E. Gabrilovich and S. Markovitch. 2007. Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In *IJCAI*.

J. Giles. 2005. Internet encyclopedias go head to head. *Nature*, 438.

A. Halevy, P. Norvig, and F. Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24.

H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis. 2010. Overview of the TAC 2010 Knowledge Base Population track. In *TAC*.

H. Ji, R. Grishman, and H. T. Dang. 2011. An overview of the TAC2011 Knowledge Base Population track. In *TAC*.

R. Koningstein, V. Spitkovsky, G. R. Harik, and N. Shazeer. 2003a. Suggesting and/or providing targeting criteria for advertisements. US Patent 2005/0228797.

R. Koningstein, V. Spitkovsky, G. R. Harik, and N. Shazeer. 2003b. Using concepts for ad targeting. US Patent 2005/0114198.

R. Koningstein, S. Lawrence, and V. Spitkovsky. 2004. Associating features with entities, such as categories of web page documents, and/or weighting such features. US Patent 2006/0149710.

P. D. Magnus. 2006. Epistemology and the Wikipedia. In *NA-CAP*.

P. McNamee and H. Dang. 2009. Overview of the TAC 2009 Knowledge Base Population track. In *TAC*.

G. A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38.

D. Milne and I. H. Witten. 2008. Learning to link with Wikipedia. In *CIKM*.

M. Recasens and M. Vila. 2010. On paraphrase and coreference. *Computational Linguistics*, 36.

V. I. Spitkovsky and A. X. Chang. 2011. Strong baselines for cross-lingual entity linking. In *TAC*.

F. M. Suchanek, G. Kasneci, and G. Weikum. 2008. YAGO: A large ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*.