

# CROSS-LANGUAGE ENTITY LINKING

PAUL MCNAMEE

JAMES MAYFIELD\*

DOUGLAS W. OARD

TAN XU

KE WU

VESELIN STOYANOV

DAVID DOERMANN



\* TODAY'S DESIGNATED BLOWHARD



# CROSS-LANGUAGE ENTITY LINKING

## KNOWLEDGE BASE

**Suzanne Collins**



Suzanne Collins at Time 100 Gala

<b>Born</b>	1962 <sup>[1]</sup> Connecticut
<b>Occupation</b>	Television scriptwriter
<b>Nationality</b>	United States
<b>Genres</b>	Fantasy Science fiction Children Young adult Suspense Action

[suzannecollinsbooks.com](http://suzannecollinsbooks.com)

**Jackie Collins**



Jackie Collins (July 2000)

<b>Born</b>	Jacqueline Jill Collins 4 October 1937 (age London, England, UK)
<b>Occupation</b>	Novelist
<b>Spouse</b>	Wallace Austin (m. 1960-1964, divorced) Oscar Lerman (m. 1966-1992, his death)
<b>Children</b>	Tracey Lerman (b. 1961)

**Susan Collins**



United States Senator  
from Maine  
Incumbent  
Assumed office

## QUERY

اطلاق النار "ألعاب الجوع" في  
ولاية كارولينا الشمالية

نحن خطوة واحدة لمشاهدة سلسلة سوزان  
تأتي في الحياة على الشاشة مع "ألعاب الجوع"  
الجوع"، وفقا لهوليوود ريبورتر.

استكمال التصوير الفوتوغرافي بشكل أساسي على  
التكليف يوزجيت في ولاية كارولينا الشمالية، حيث تم  
اطلاق النار على الفيلم، يوم السبت الماضي.

"ألعاب الجوع" وكان من المقرر في البداية  
باعتباره نفض الغبار من تصف الميزانية للشباب،  
ولكن نظرا لشعبية واسعة النطاق من سلسلة كتاب،  
أصبح من إنتاج ما يقرب من 100 مليون دولار، وتقارير  
صحيفة لوس أنجلوس تايمز. يوزجيت تستهدف رواد  
السينما من جميع الأعمار، وليس فقط المراهقين.

ثلاثية وأكثر من 12 مليون نسخة في الطباعة.



# THE UNIVERSAL HAMMER



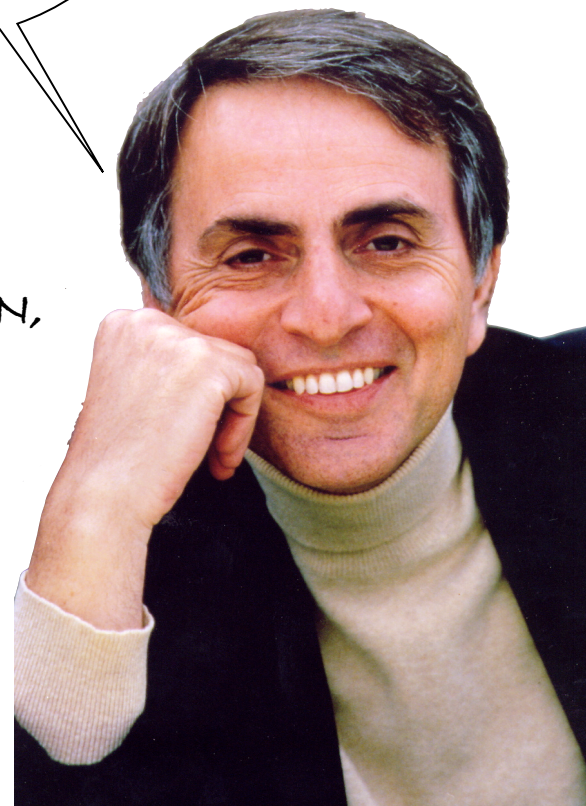


# FEATURES

---

- ☐ NAME-MATCHING
  - ☐ ACRONYMS, ALIASES, STRING SIMILARITY
- ☐ DOCUMENT FEATURES
  - ☐ TF/IDF-WEIGHTED COMPARISONS, OCCURRENCE OF KB FACTS IN QUERY TEXT
- ☐ ENTITY TYPE, NAMED ENTITY CO-OCCURRENCES
  - ☐ TYPE (I.E., IS THIS A PERSON, ORGANIZATION, LOCATION?)
  - ☐ DO OTHER ENTITIES CO-OCCUR IN QUERY DOCUMENT AND KB RECORD?
- ☐ ABSENCE (NIL INDICATIONS)
  - ☐ DOES ANY CANDIDATE LOOK LIKE A VIABLE MATCH?

BILLIONS AND  
BILLIONS OF  
FEATURES

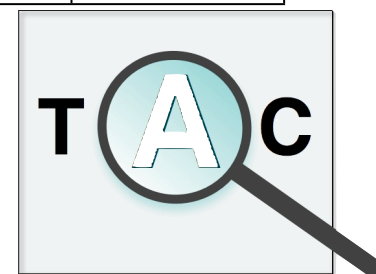






# TAC 2011 MONOLINGUAL RESULTS

Run	Description	2010 Micro	B <sup>3</sup> Precision	B <sup>3</sup> Recall	B <sup>3</sup> F1
<b>English Entity Linking</b>					
hltcoe1	Exact name match	0.772	0.730	0.750	0.740
hltcoe2	Supervised classification	0.772	0.724	0.748	0.736
hltcoe3	Augmented KB, Supervised classification	0.728	0.681	0.701	0.691
<b>English Entity Linking - No Wiki</b>					
hltcoe1	Exact name match	0.749	0.707	0.720	<b>0.714</b>
hltcoe2	Supervised classification	0.749	0.702	0.717	0.710





# WE WERE BORED



- ☐ LET'S BUILD A NEW CROSS-LANGUAGE ENTITY LINKING COLLECTION!
- ☐ BUT WHAT LANGUAGE?
- ☐ TWENTY-ONE LANGUAGES
- ☐ OVER 55,000 QUERIES
- ☐ PUBLICLY AVAILABLE AT [HLTCOE.ORG/DATASETS](http://HLTCOE.ORG/DATASETS)

**ALBANIAN (SQ)**  
**ARABIC (AR)**  
**BULGARIAN (BG)**  
**CHINESE (ZH)**  
**CROATIAN (HR)**  
**CZECH (CS)**

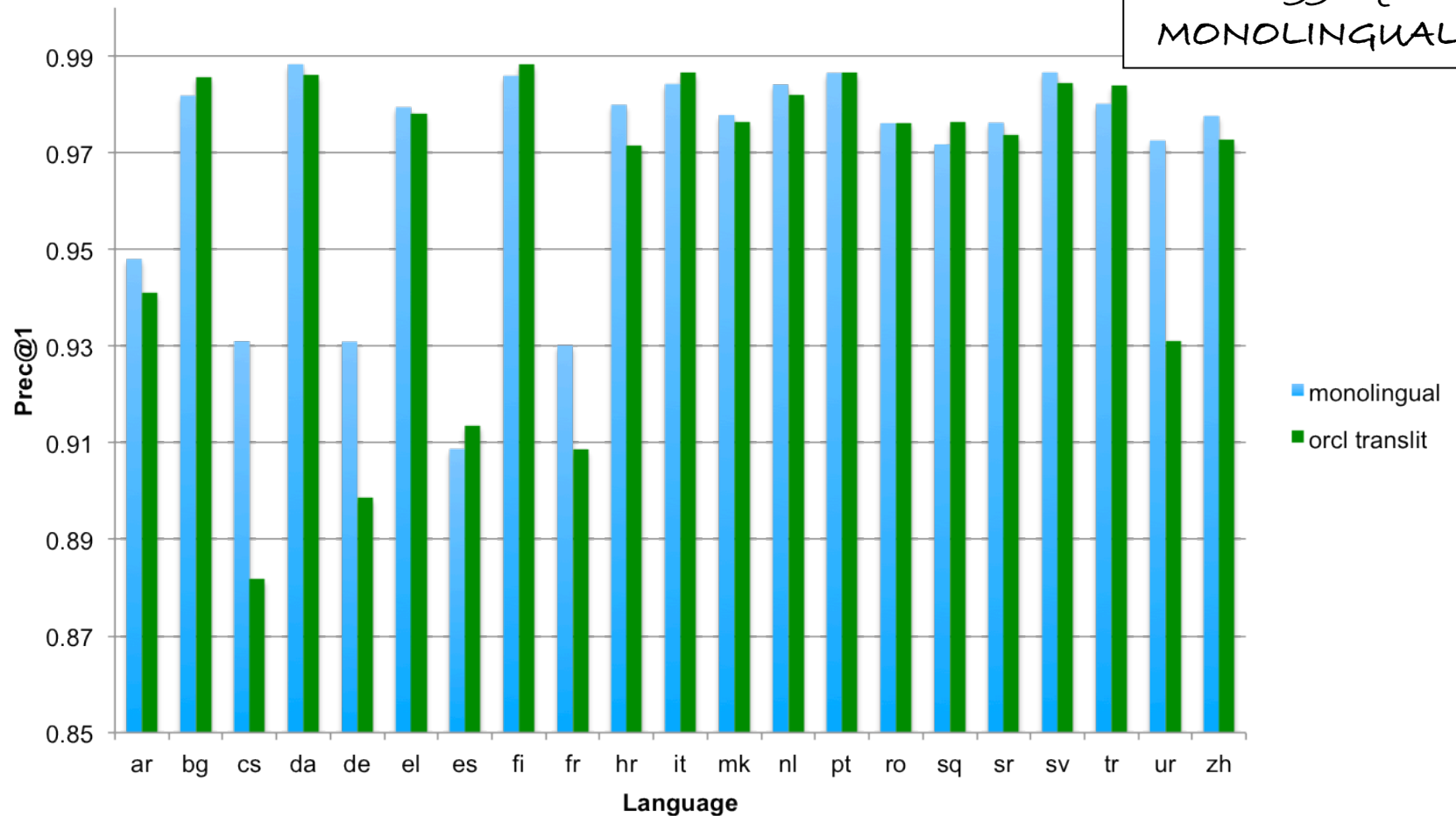
**DANISH (DA)**  
**DUTCH (NL)**  
**FINNISH (FI)**  
**FRENCH (FR)**  
**GERMAN (DE)**  
**GREEK (EL)**  
**ITALIAN (IT)**  
**MACEDONIAN (MK)**  
**PORTUGUESE (PT)**  
**ROMANIAN (RO)**  
**SERBIAN (SR)**  
**SPANISH (ES)**  
**SWEDISH (SV)**  
**TURKISH (TR)**  
**URDU (UR)**



# PERFECT TRANSLITERATION

## English vs. Perfect Transliteration

MEAN: 99.2% OF  
MONOLINGUAL

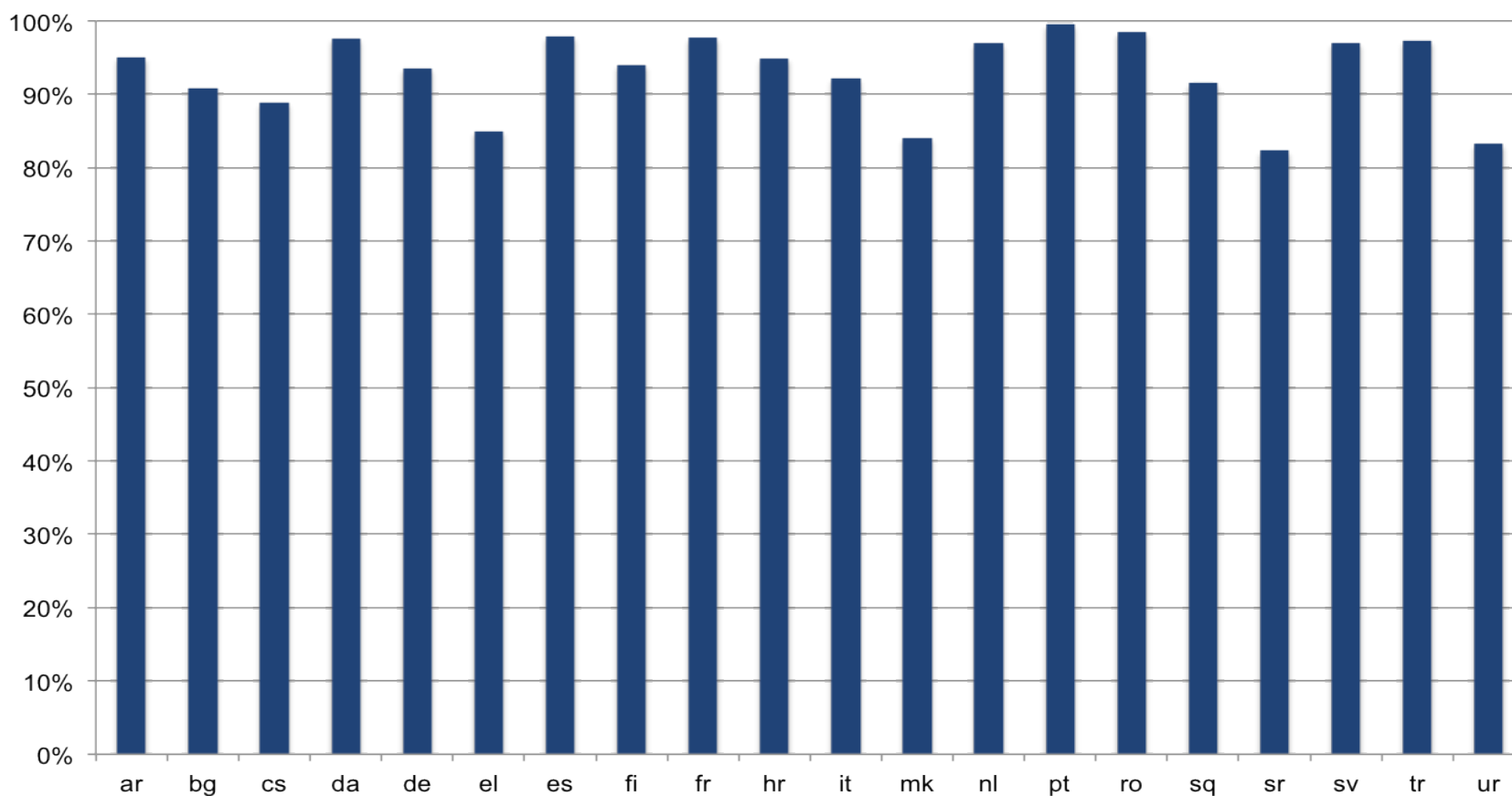




# STATISTICAL TRANSLITERATION

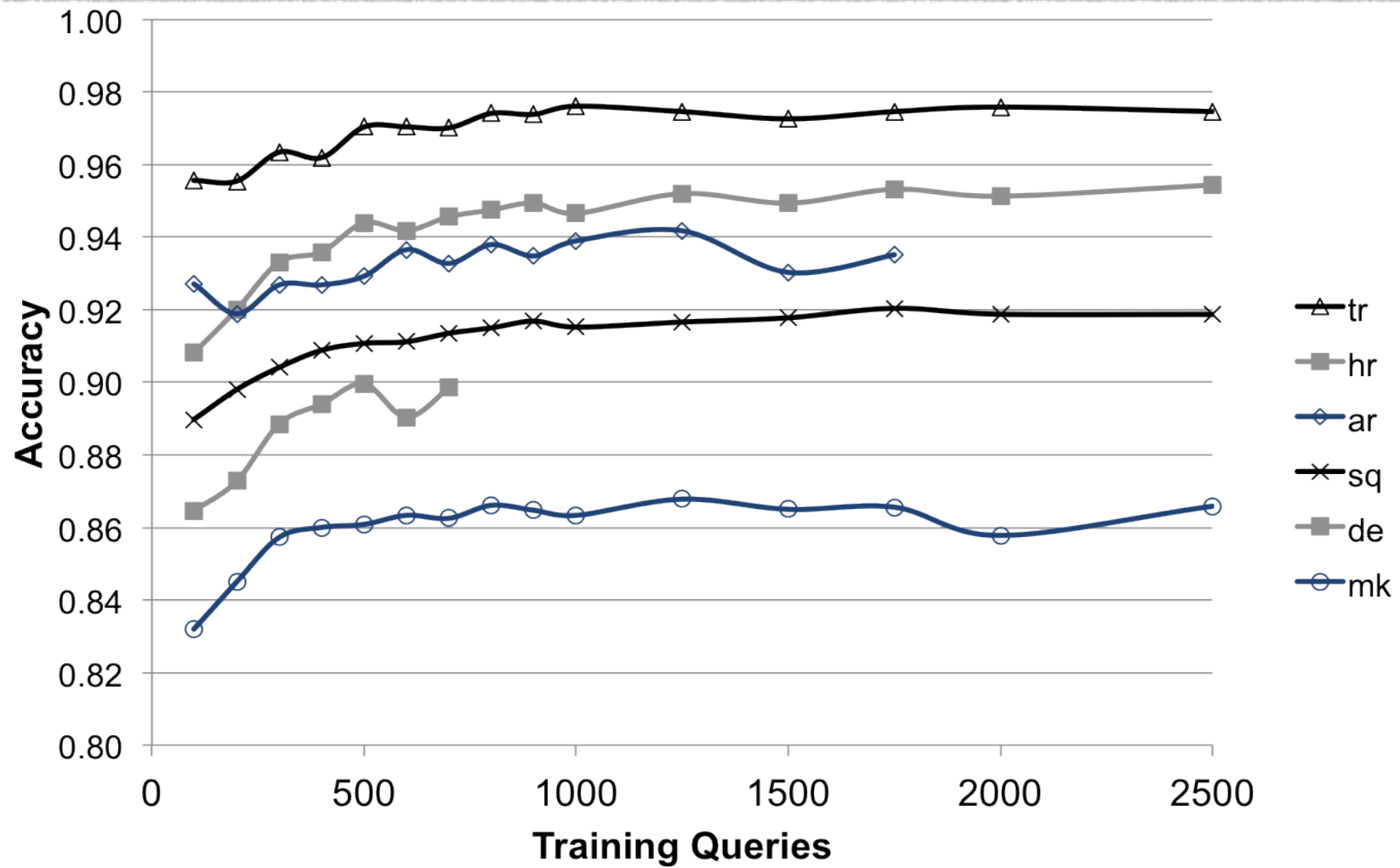
MEAN: 93% OF  
MONOLINGUAL

**% Monolingual Performance**





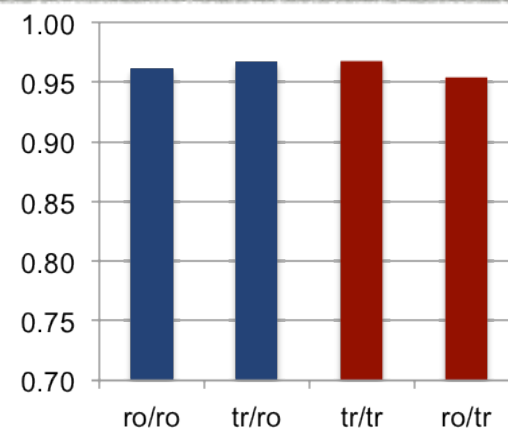
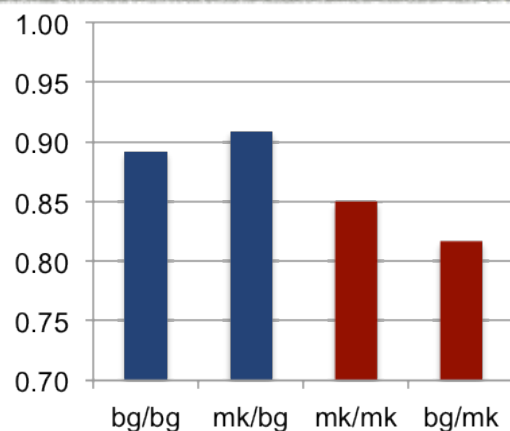
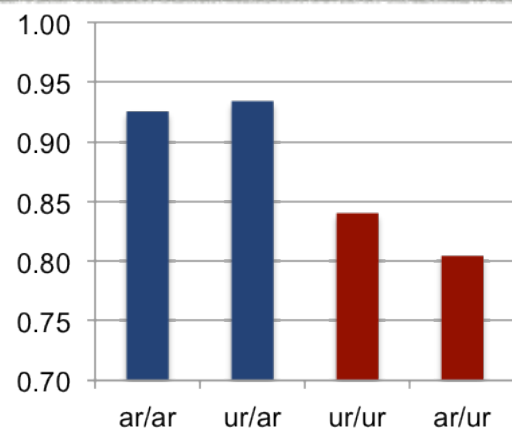
# HOW MUCH TRAINING DATA IS REQUIRED?







# AN ASIDE: OFF-LANGUAGE TRAINING



- THREE LANGUAGE PAIRS USING SAME WRITING SYSTEM
  - ARABIC/URDU
  - BULGARIAN/MACEDONIAN
  - ROMANIAN/TURKISH
- CAN TRAINING ON A DIFFERENT LANGUAGE BE EFFECTIVE?
- LITTLE DEGRADATION OBSERVED; FEATURES APPEAR TO BE LARGELY LANGUAGE AGNOSTIC



# THE UNIVERSAL HAMMER

---





# TAC 2011 CROSS-LANGUAGE ENTITY LINKING


---

☐ THREE MAIN COMPONENTS (OTHER THAN  ):

1. NAME MAPPING

2. CONTEXT MAPPING

 CROSS-LANGUAGE IR

 MACHINE TRANSLATION

3. NIL CLUSTERING

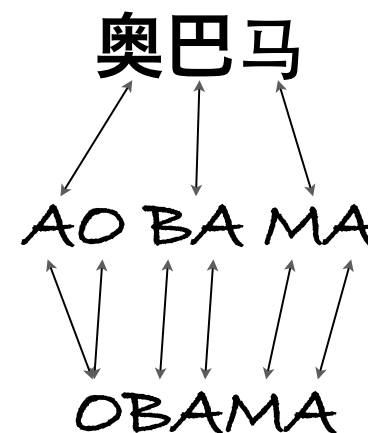




# UNIVERSAL TRANSLITERATION

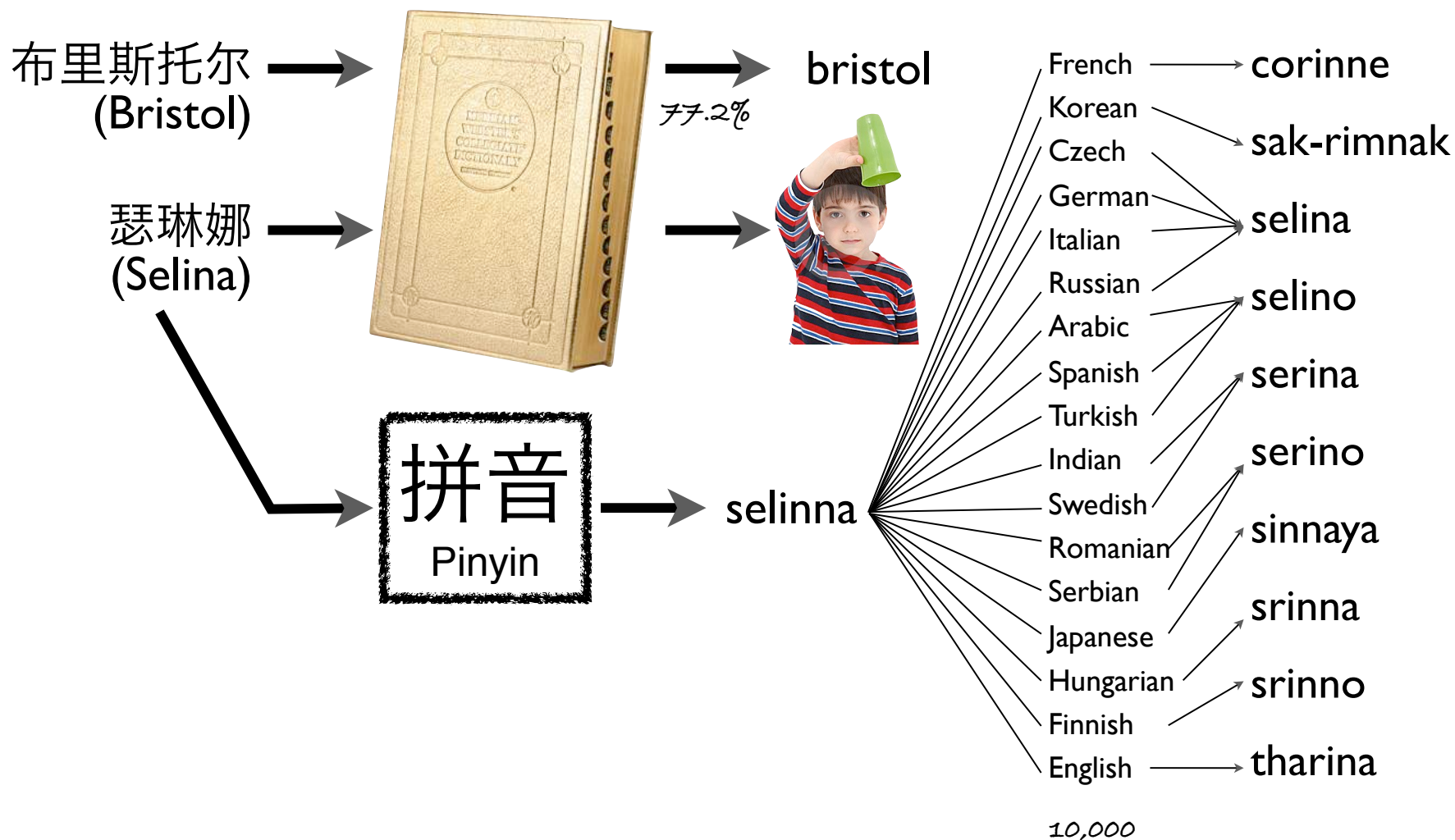


- ☐ IRVINE ET AL. 2010: TREAT TRANSLITERATION AS PHRASE-BASED STATISTICAL MACHINE TRANSLATION
- ☐ CHARACTERS ARE THE 'WORDS' TO BE TRANSLATED
- ☐ TRAINING "SENTENCES" ARE NAME PAIRS EXTRACTED FROM WIKIPEDIA CROSS-LANGUAGE ARTICLE TITLES
- ☐ FOR CHINESE/ENGLISH TRANSLITERATION, WE USE THIS APPROACH TO MAP FROM PINYIN TO ENGLISH





# CHINESE-ENGLISH TRANSLITERATION







# CHINESE TRANSLITERATION RESOURCES

---

<i>Dictionary Name</i>	<i>Source</i>	<i>Size</i>
Names of the World's Peoples (Guo 2007)	Xinhua News Agency	676,871
Place Names of the World (Zhou 2008)	Xinhua News Agency	177,372
Chinese English Name Entity Lists v1.0 (Huang 2005)	LDC (LDC2005T34)	122,344
Chinese English Cross-Lingual Name Pairs	Chinese Wikipedia	427,678





# TAC 2011 CROSS-LANGUAGE ENTITY LINKING

---

□ THREE MAIN COMPONENTS (OTHER THAN  ):

1. NAME MAPPING

2. CONTEXT MAPPING

 CROSS-LANGUAGE IR

 MACHINE TRANSLATION

3. NIL CLUSTERING





# MACHINE TRANSLATION: ORIGINAL DOCUMENT

《星球大战》道具将随发现号航天飞机上天

2007年10月11日09:40

美国“发现”号航天飞机计划本月23日发射升空。届时,著名科幻电影《星球大战》中主人公“天行者卢克”的武器“光剑”将随飞机一同飞赴太空,完成一次太空之旅,以纪念“星战”系列电影问世30周年。

“光剑”将上天

美国著名导演乔治·卢卡斯1977年推出首部《星球大战》影片大获成功后,又陆续完成5部续集影片。这组著名“星战”系列电影,成为史上最成功的科幻电影之一。

这次将搭乘“发现”号上天的“光剑”,正是首部“星战”影片拍摄时天行者卢克的扮演者马克·哈米尔使用的道具。

据英国《每日邮报》9日报道,这把形似激光光线的“光剑”由导演卢卡斯亲自乘飞机送抵美国国家航空和航天局(NASA)位于得克萨斯州休斯敦的约翰逊航天中心。奔赴太空之前,“光剑”在中心专供游客参观的“休斯敦航天中心”向公众展出。

随卢卡斯护送“光剑”的还有由一些演员装扮的“星战”人物,包括体型高大、身披毛发的乌奇族战士“楚巴”等。

“发现”号7名机组成员9日在位于佛罗里达州的肯尼迪航天中心完成起飞前最后一次演练。

“光剑”则将被装入“发现”号机舱内,与宇航员们一同飞赴太空,完成一次为期13天的太空之旅。

30周年纪念

此次“光剑”上天是“星战”电影问世30周年纪念活动的一部分。“休斯敦航天中心”工作人员道格·马蒂斯:“《星球大战》电影和航天飞机在美国和世界文化中都深入人心,因此与天行者的‘光剑’一同飞天是合适的纪念方式”。

虽然NASA坚称“光剑”在“旅行”期间将始终置于“发现”号机舱内,但是指挥中心工作人员说,宇航员们在完成为太空站安装新太空舱等重要任务后,也许会有兴致把“光剑”带到机舱外“比试”一番



# MT:TRANSLATED DOCUMENT

《 star wars 》 props space shuttle discovery to heaven

2007 year october 11 , 2007 09:40

the united states " found space shuttle program launched on the 23rd this month .

at that time , 著名 science fiction movie 《 》 star wars in the owner of the public " anakin skywalker luke " weapons " light sword " will together with the aircraft to fly to space , completing a space journey , to commemorate " lightsaber " 30th anniversary of the movie series appeared .

" light sword " to heaven

us 著名 director <<<3 乔治 · 卢卡斯 >>>3 1977 launched last year 's first 《 star wars 》 film after great success , and completed five films , sequel .

this group of 著名 " lightsaber " movie series , has become one of the most successful science fiction movie history .

this will take " " found space " light sword " , is the first " lightsaber " film when anakin skywalker luke the 马克 · 哈米尔 use the props .

according to the british 《 》 post daily reported on the 9th , the scale of laser light " 光剑 " personally by director

<<<22 lucas >>>22 left reach the us space agency nasa ( nasa ) the johnson space center in houston , texas .

prior to fly into space , " 光剑 " " houston space center " in the center offers tourists visited " to the public display .

with <<<1 lucas >>>1 " sword also " alone were escorted by some actors dressed up as the " lightsaber " figures , including body tall 、 wearing her hair the 乌奇族 soldiers " 楚巴卡 " .

" found " no. 7 crew members on the 9th in the completion of the kennedy space center in florida last drill before takeoff .

" light sword " was found to be encased in " " no. inside the cabin , together with astronauts to space , completing a





# MT: RE-CAPITALIZE

《 Star Wars 》 props space Shuttle Discovery to Heaven

2007 year October 11 , 2007 09:40

the United States " found space Shuttle program launched on the 23rd this month .

at that time , 著名 Science Fiction movie 《 》 Star Wars in the owner of the public " Anakin Skywalker Luke "

weapons " light sword " will together with the aircraft to fly to space , completing a space journey , to commemorate " lightsaber " 30th Anniversary of the movie Series appeared .

" light sword " to Heaven

US 著名 Director <<< 3 乔治 · 卢卡斯 >>> 1977 launched last year 's first 《 Star Wars 》 film after great success ,

and completed five films , sequel .

this Group of 著名 " lightsaber " movie Series , has become one of the most successful Science Fiction movie

History .

this will take " " found space " light sword " , is the first " lightsaber " film when Anakin Skywalker Luke the 马克 · 哈米

尔 use the props .

according to the British 《 》 Post Daily reported on the 9th , the scale of Laser light " 光剑 " personally by Director

<<< 22 Lucas >>> 22 left reach the US space agency NASA NASA the Johnson space Center in Houston , Texas .  
prior to fly into space , " 光剑 " " Houston space Center in the Center offers tourists visited " to the public display .

with <<< 1 Lucas >>> 1 " sword also " alone were escorted by some actors dressed up as the " lightsaber " figures ,  
including body tall 、 wearing her hair the 乌奇族 soldiers 楚巴卡 .

" found " No. 7 crew members on the 9th in the completion of the Kennedy space Center in Florida last drill before  
takeoff .



# MT: TRANSLITERATE REMAINING CHINESE CHARACTERS

《 Star Wars 》 props space Shuttle Discovery to Heaven

2007 year October 11 , 2007 09:40

the United States “ ” found space Shuttle program launched on the 28th of October at that time , Zhuming Science Fiction movie 《 》 Star Wars in the “ Skywalker Luke ”

weapons “ light sword ” will together with the aircraft to fly to space to commemorate “ lightsaber ” 30th Anniversary of the movie Series appeared .

“ light sword ” to Heaven

US Zhuming Director <<<3 George Lucas >>>3 1977 launched the Star Wars film after great

success , and completed five films , sequel .

this Group of Zhuming “ lightsaber ” movie Series , has become one of the most successful science Fiction movie History .

this will take “ ” found space “ light sword ” , is the first “ lightsaber ” film making Skywalker Luke the Mark Hamill use the props .

according to the British 《 》 Post Daily reported on the 9th , the special “ light sword ” personally by

Director <<<22 Lucas >>>22 left reach the US space agency NASA Johnson Space Center in Houston , Texas .

prior to fly into space , “ lightsaber ” “ Houston space Center in the Center offers tourists “ ” to the public display .

with <<<1 Lucas >>>1 “ sword also ” alone were escorted by “ ” actors dressed up as the lightsaber figures , including body tall 、 wearing her hair the wuqizu soldiers tshubaka .

“ found ” No. 7 crew members on the 9th in the completion of the Kennedy space Center in Florida last drill before takeoff .



**WOOKIE SOLDIER CHEWBACCA**



# TAC 2011 CROSS-LANGUAGE ENTITY LINKING

---

□ THREE MAIN COMPONENTS (OTHER THAN  ):

1. NAME MAPPING

2. CONTEXT MAPPING

 CROSS-LANGUAGE IR

 MACHINE TRANSLATION

3. NIL CLUSTERING





# NIL CLUSTERING

---



- ☐ BUILD A CLASSIFIER TO DETERMINE WHETHER TWO NIL MENTIONS ARE COREFERENT
  - ☐ USE THE SAME FEATURES AS USED IN ENTITY LINKING
- ☐ EACH RESULTING CONNECTED COMPONENT IS CONSIDERED TO BE A UNIQUE ENTITY
- ☐ STRAWMAN SYSTEM TO MAKE OUR APPROACH LOOK GOOD: EXACT STRING MATCH
- ☐ RESULTS:
  - ☐ CLASSIFIER: 70.5%
  - ☐ STRAWMAN: 72.5%





# TAC 2011 CROSS-LANGUAGE RESULTS

- ☐ ALL RUNS USED
  - ☐ CLIR
  - ☐ MACHINE TRANSLATION
  - ☐ ENGLISH NAMED ENTITY RECOGNITION



Run	Description	2010 Micro	B <sup>3</sup> Precision	B <sup>3</sup> Recall	B <sup>3</sup> F1
hltcoe1	Exact name match	0.800	0.702	0.779	0.738
hltcoe2	Supervised classification, multiple transliterations	0.790	0.708	0.703	0.705
hltcoe3	Exact name match, multiple transliterations	0.790	0.689	0.765	0.725



# CONCLUSIONS

---

- ☐ NAME MAPPING IS CRUCIAL FOR CROSS-LANGUAGE ENTITY LINKING
- ☐ THE MANY-FEATURED MACHINE LEARNING APPROACH TO MONOLINGUAL ENTITY LINKING WORKS WELL FOR TRANSLINGUAL ENTITY LINKING TOO
- ☐ REVEALING CONCLUSIONS ONE AT A TIME IS ANNOYING
- ☐ NIL CLUSTERING BY EXACT NAME MATCHING WORKS DISTURBINGLY WELL
- ☐ XLEL-21 COLLECTION NOW AVAILABLE FROM [HLTCOE.ORG/DATASETS](http://HLTCOE.ORG/DATASETS)



GRATUITOUS  
IMAGE

SPECIAL THANKS TO:  
CHRIS CALLISON-BURCH  
VLAD EIDELMAN  
KRISTY HOLLINGSHEAD  
ANN IRVINE  
DAWN LAWRIE  
SCOTT ROBERTS  
BEN SHAYNE





THANK YOU