Mining and Modeling Relations between Formal and Informal Chinese Phrases from Web Corpora

Zhifei Li and David Yarowsky
Johns Hopkins University

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B:

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-C:

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-C: 88

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-C: 88

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-C: 88 ???

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-C: 88 ???

Language from Mars?

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-C: 88 ???

Language from Mars?

88 = bye-bye = 拜拜

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-C: 88 ???

Language from Mars?

88 = *bye-bye* = 拜拜

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-C: 88 ???

Language from Mars?

88 = *bye-bye* = 拜拜

BaBa *BaiBai*

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-C: 88 ???

Language from Mars?

88 = bye-bye = 拜拜

BaBa *BaiBai*

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-C: 88 ???

Language from Mars?

88 = bye-bye = 拜拜

BaBa *BaiBai*

- **Our task: mining *formal-informal* relations from web**

Person-A: 对不起， 我要先下线了，拜拜

Translation: *Sorry, i got to go, bye bye*

Person-B: bye bye

Person-C: 88 ???

Language from Mars?

88 = bye-bye = 拜拜

BaBa *BaiBai*

- **Our task: mining *formal-informal* relations from web**
  - Identify *informal* phrases
  - Find their *formal* equivalents

Why is this task important?

- **Enormous amount of informal text on the web**
  - Online-chat, Bulletin Board System (BBS), Web Blog, and so on
- **New informal phrases are created everyday**
- **A general text-processing system does not work well on informal text**
  - machine translation
  - information retrieval
- **Our method can be used for text normalization**

Outline

- **Rule-driven intuition**
- **Data-driven intuition**
- **Our approach**
  - a log-linear model combining both rule- and data-driven intuitions
- **Experimental results**

Data Collection

- **We manually collect about 900 examples**
  - used for both training and testing using cross-validation
- **We then classify the examples into different classes based on our knowledge in Chinese and linguistics**

Homophones

Homophones

- **Similar pronunciation but different meaning or written-form**

Homophones

- **Similar pronunciation but different meaning or written-form**

Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%

Homophones

- **Similar pronunciation but different meaning or written-form**

Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%

Homophones

- **Similar pronunciation but different meaning or written-form**

Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%

Homophones

- **Similar pronunciation but different meaning or written-form**

Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%

Homophones

- **Similar pronunciation but different meaning or written-form**

Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%

Homophones

- **Similar pronunciation but different meaning or written-form**

Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%

Homophones

- **Similar pronunciation but different meaning or written-form**

Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%

Homophones

- **Similar pronunciation but different meaning or written-form**

Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%

Homophones

- **Similar pronunciation but different meaning or written-form**

Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%

Formal

Homophones

- **Similar pronunciation but different meaning or written-form**

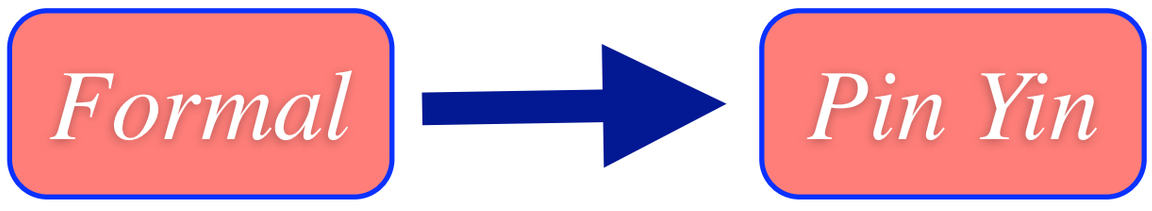
Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%

Formal →

Homophones

- **Similar pronunciation but different meaning or written-form**

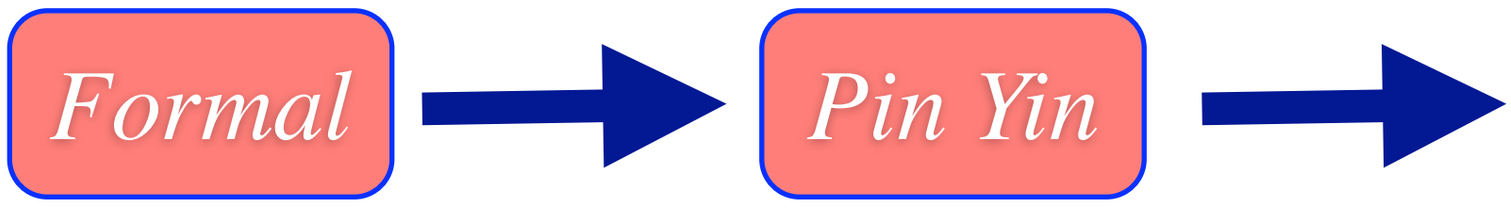
Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%



Homophones

- **Similar pronunciation but different meaning or written-form**

Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%



Homophones

- **Similar pronunciation but different meaning or written-form**

Formal	Informal	Percentage
版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4%
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4%
拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21%

Abbreviations and Acronyms

Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (N$\ddot{u}$PengYou)[girl friend]	GF	7%

Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (N$\ddot{u}$PengYou)[girl friend]	GF	7%

Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (N$\ddot{u}$PengYou)[girl friend]	GF	7%

Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (N*ü*PengYou)[girl friend]	GF	7%

Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (N*ü*PengYou)[girl friend]	GF	7%

Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (N*ü*PengYou)[girl friend]	GF	7%

Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (N*ü*PengYou)[girl friend]	GF	7%

Formal

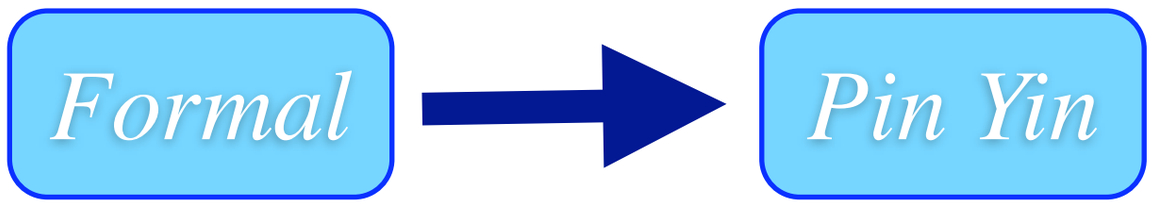
Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (N*ü*PengYou)[girl friend]	GF	7%

Formal →

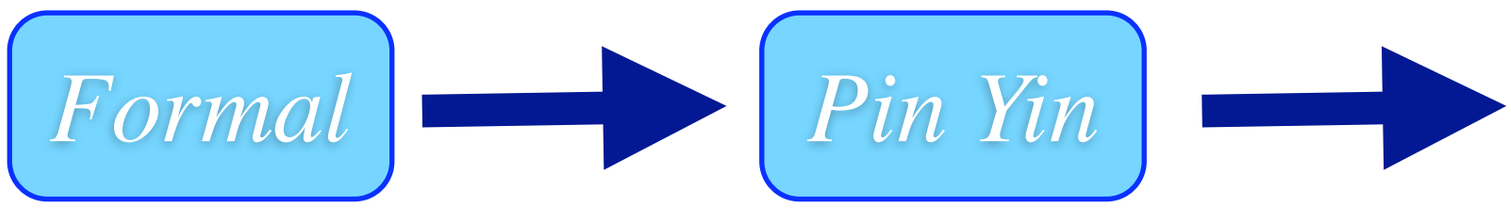
Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (NüPengYou)[girl friend]	GF	7%



Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (NüPengYou)[girl friend]	GF	7%

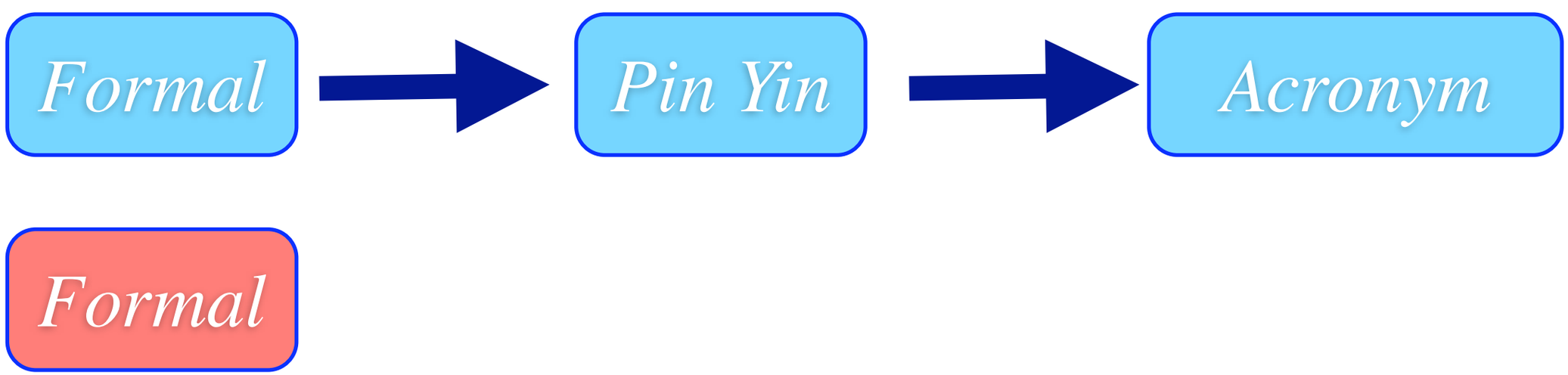


Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (N*ü*PengYou)[girl friend]	GF	7%

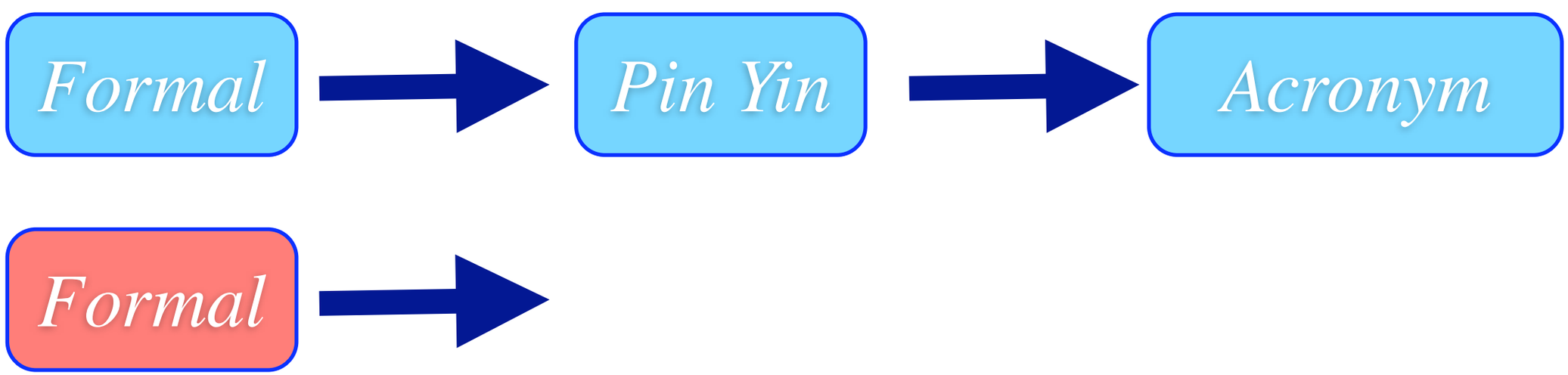
Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (NüPengYou)[girl friend]	GF	7%



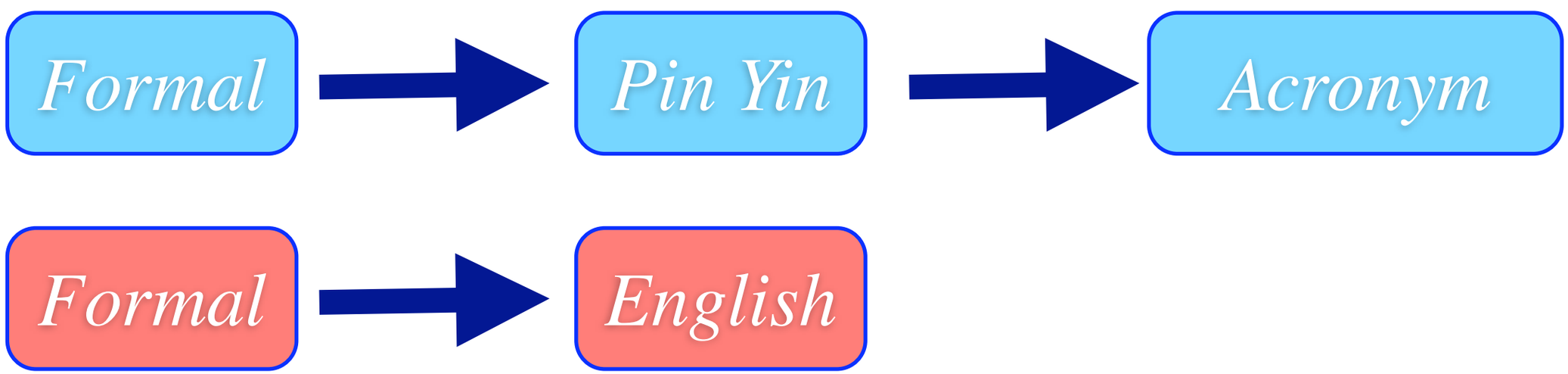
Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (NüPengYou)[girl friend]	GF	7%



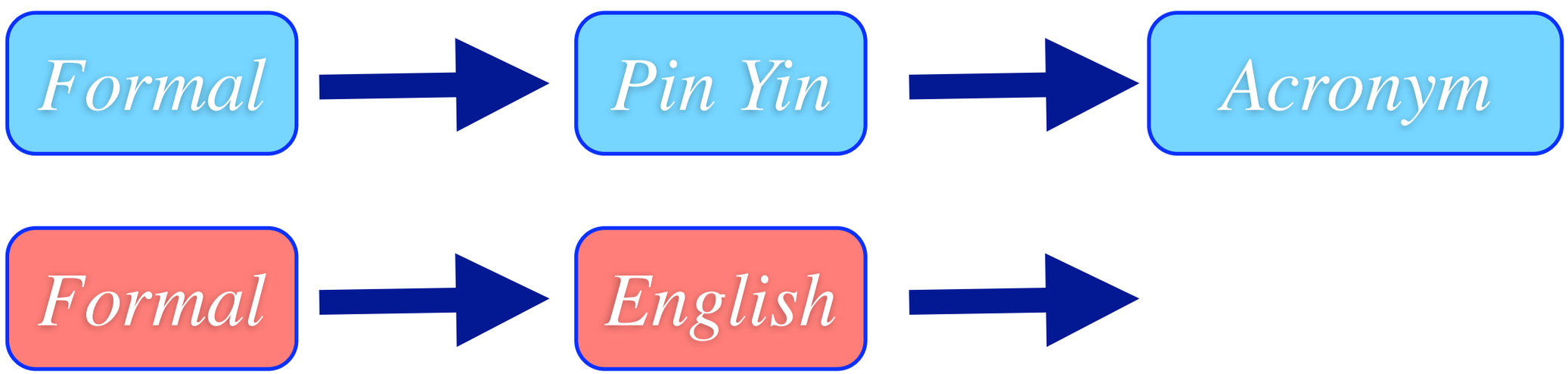
Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (NüPengYou)[girl friend]	GF	7%



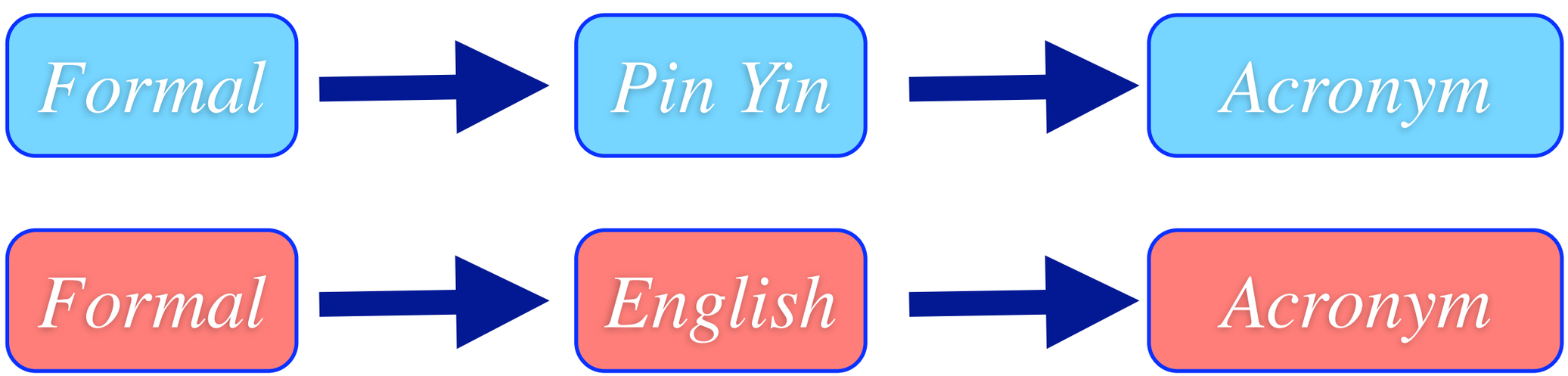
Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (NüPengYou)[girl friend]	GF	7%



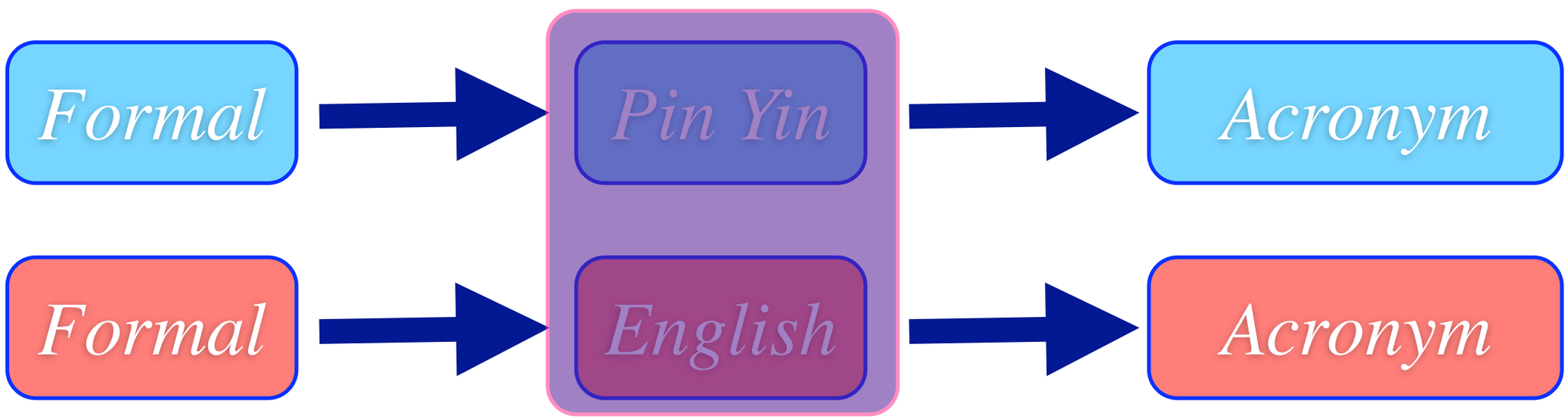
Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (NüPengYou)[girl friend]	GF	7%



Abbreviations and Acronyms

Formal	Informal	Percentage
美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	4%
哥哥 (GeGe)[elder brother]	GG	12%
女朋友 (NüPengYou)[girl friend]	GF	7%



Transliterations

Transliterations

Formal	Informal	Percentage
歌迷 (GeMi)[fans]	粉丝 (FenSi)[a chinese food]	2 %
谢谢 (XieXie)[thank you]	3Q (SanQiu)	

Transliterations

Formal	Informal	Percentage
歌迷 (GeMi)[fans]	粉丝 (FenSi)[a chinese food]	2 %
谢谢 (XieXie)[thank you]	3Q (SanQiu)	

Transliterations

Formal	Informal	Percentage
歌迷 (GeMi)[fans]	粉丝 (FenSi)[a chinese food]	2 %
谢谢 (XieXie)[thank you]	3Q (SanQiu)	

Transliterations

Formal	Informal	Percentage
歌迷 (GeMi)[fans]	粉丝 (FenSi)[a chinese food]	2 %
谢谢 (XieXie)[thank you]	3Q (SanQiu)	

Transliterations

Formal	Informal	Percentage
歌迷 (GeMi)[fans]	粉丝 (FenSi)[a chinese food]	2 %
谢谢 (XieXie)[thank you]	3Q (SanQiu)	

Transliterations

Formal	Informal	Percentage
歌迷 (GeMi)[fans]	粉丝 (FenSi)[a chinese food]	2 %
谢谢 (XieXie)[thank you]	3Q (SanQiu)	

Formal

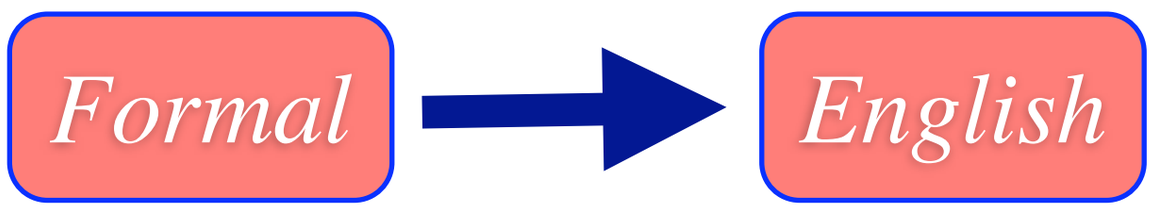
Transliterations

Formal	Informal	Percentage
歌迷 (GeMi)[fans]	粉丝 (FenSi)[a chinese food]	2 %
谢谢 (XieXie)[thank you]	3Q (SanQiu)	

Formal →

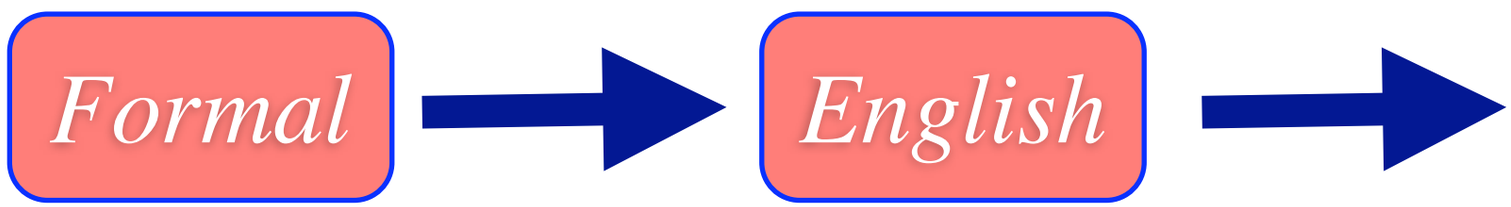
Transliterations

Formal	Informal	Percentage
歌迷 (GeMi)[fans]	粉丝 (FenSi)[a chinese food]	2 %
谢谢 (XieXie)[thank you]	3Q (SanQiu)	



Transliterations

Formal	Informal	Percentage
歌迷 (GeMi)[fans]	粉丝 (FenSi)[a chinese food]	2 %
谢谢 (XieXie)[thank you]	3Q (SanQiu)	



Transliterations

Formal	Informal	Percentage
歌迷 (GeMi)[fans]	粉丝 (FenSi)[a chinese food]	2 %
谢谢 (XieXie)[thank you]	3Q (SanQiu)	

Other phenomena

Other phenomena

Formal	Informal	Percentage
希拉里粉丝 (XiLaLiFenSi)[fans of Hilary]	稀饭 (XiFan)[gruel]	45%
奥巴马粉丝 (AoBaMaFenSi)[fans of Obama]	奥饭 (AoFan)	
超强 (ChaoQiang)[super strong]	走召弓虽 (ZouZhaoGongXu)	

Other phenomena

Formal	Informal	Percentage
希拉里粉丝 (XiLaLiFenSi)[fans of Hilary]	稀饭 (XiFan)[gruel]	45%
奥巴马粉丝 (AoBaMaFenSi)[fans of Obama]	奥饭 (AoFan)	
超强 (ChaoQiang)[super strong]	走召弓虽 (ZouZhaoGongXu)	

Other phenomena

Formal	Informal	Percentage
希拉里粉丝 (XiLaLiFenSi)[fans of Hilary]	稀饭 (XiFan)[gruel]	45%
奥巴马粉丝 (AoBaMaFenSi)[fans of Obama]	奥饭 (AoFan)	
超强 (ChaoQiang)[super strong]	走召弓虽 (ZouZhaoGongXu)	

Other phenomena

Formal	Informal	Percentage
希拉里粉丝 (XiLaLiFenSi)[fans of Hilary]	稀饭 (XiFan)[gruel]	45%
奥巴马粉丝 (AoBaMaFenSi)[fans of Obama]	奥饭 (AoFan)	
超强 (ChaoQiang)[super strong]	走召弓虽 (ZouZhaoGongXu)	

Other phenomena

Formal	Informal	Percentage
希拉里粉丝 (XiLaLiFenSi)[fans of Hilary]	稀饭 (XiFan)[gruel]	45%
奥巴马粉丝 (AoBaMaFenSi)[fans of Obama]	奥饭 (AoFan)	
超强 (ChaoQiang)[super strong]	走召弓虽 (ZouZhaoGongXu)	

超强

Other phenomena

Formal	Informal	Percentage
希拉里粉丝 (XiLaLiFenSi)[fans of Hilary]	稀饭 (XiFan)[gruel]	45%
奥巴马粉丝 (AoBaMaFenSi)[fans of Obama]	奥饭 (AoFan)	
超强 (ChaoQiang)[super strong]	走召弓虽 (ZouZhaoGongXu)	

超强

走 召

Other phenomena

Formal	Informal	Percentage
希拉里粉丝 (XiLaLiFenSi)[fans of Hilary]	稀饭 (XiFan)[gruel]	45%
奥巴马粉丝 (AoBaMaFenSi)[fans of Obama]	奥饭 (AoFan)	
超强 (ChaoQiang)[super strong]	走召弓虽 (ZouZhaoGongXu)	

超强

走 召 弓 虽

Other phenomena

Formal	Informal	Percentage
希拉里粉丝 (XiLaLiFenSi)[fans of Hilary]	稀饭 (XiFan)[gruel]	45%
奥巴马粉丝 (AoBaMaFenSi)[fans of Obama]	奥饭 (AoFan)	
超强 (ChaoQiang)[super strong]	走召弓虽 (ZouZhaoGongXu)	

超强

走 召 弓 虽

Web users are very very creative!

Definitions in Real Text

- **In practice, the corresponding formal equivalent can sometimes be found near the informal phrases**
  - direct definitions
  - indirect associations

Direct Definitions

Definition Text	Relation
GF是女朋友的意思。	(女朋友, GF)
香港的食水采用世界卫生组织 (世卫)饮用水水质指引……	(世界卫生组织, 世卫)

Direct Definitions

Definition Text	Relation
GF是女朋友的意思。	(女朋友, GF)
香港的食水采用世界卫生组织 (世卫)饮用水水质指引……	(世界卫生组织, 世卫)

Direct Definitions

Definition Text	Relation
GF是女朋友的意思。	(女朋友, GF)
香港的食水采用世界卫生组织 (世卫)饮用水水质指引……	(世界卫生组织, 世卫)

Direct Definitions

Definition Text	Relation
GF是女朋友的意思。	(女朋友, GF)
香港的食水采用世界卫生组织 (世卫)饮用水水质指引……	(世界卫生组织, 世卫)

Pattern: informal是formal 的意思

Direct Definitions

Definition Text	Relation
GF是女朋友的意思。	(女朋友, GF)
香港的食水采用世界卫生组织 (世卫)饮用水水质指引……	(世界卫生组织, 世卫)

Pattern: informal是formal 的意思

Direct Definitions

Definition Text	Relation
GF是女朋友的意思。	(女朋友, GF)
香港的食水采用世界卫生组织 (世卫)饮用水水质指引……	(世界卫生组织, 世卫)

Pattern: informal是formal 的意思

Direct Definitions

Definition Text	Relation
GF是女朋友的意思。	(女朋友, GF)
香港的食水采用世界卫生组织(世卫)饮用水水质指引……	(世界卫生组织, 世卫)

Pattern: informal是formal 的意思

Direct Definitions

Definition Text	Relation
GF是女朋友的意思。	(女朋友, GF)
香港的食水采用世界卫生组织 (世卫)饮用水水质指引……	(世界卫生组织, 世卫)

Pattern: *informal*是*formal* 的意思

Pattern: *formal*(*informal*)

Indirect Associations in Online Chat

⋮

Person A: 对不起， 我要先下线了, 拜拜

Person B: 88

Indirect Associations in Online Chat

⋮

Person A: 对不起， 我要先下线了,拜拜

Person B: 88

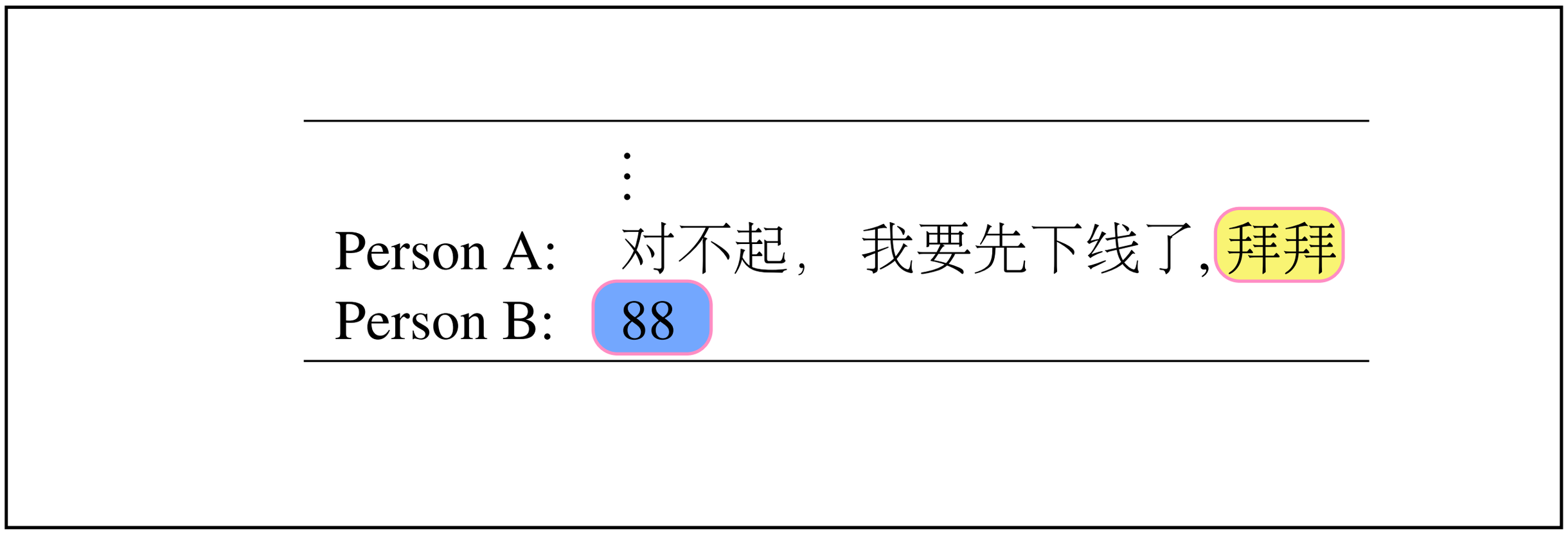
Indirect Associations in Online Chat

⋮

Person A: 对不起， 我要先下线了，拜拜

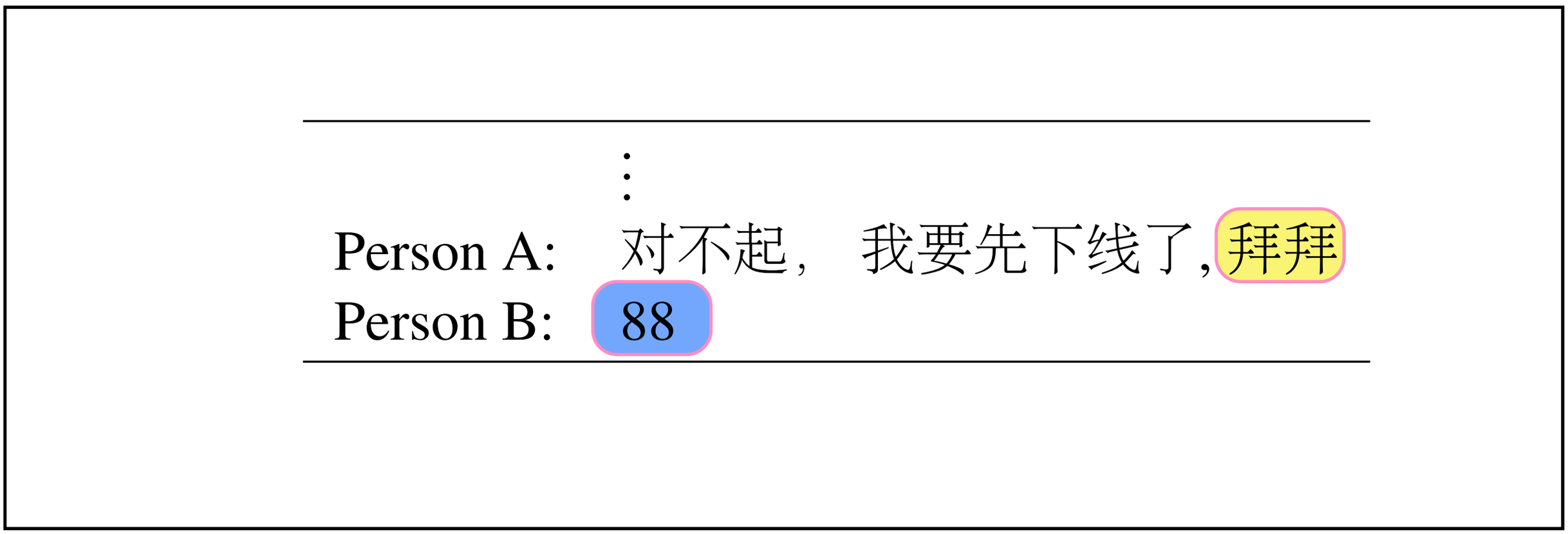
Person B: 88

Indirect Associations in Online Chat



(拜拜， 88)

Indirect Associations in Online Chat



(拜拜， 88)

Bye bye

Indirect Associations in News Articles

Indirect Associations in News Articles

Title	都灵**冬奥会**开幕式将激情上演
Text	新华社都灵2月9日电(记者丁莹阎涛)第20届**冬季奥运会**的开幕式将于当地时间10日晚8点在都灵奥林匹克体育场正式揭开神秘的面纱。

Indirect Associations in News Articles

Title	都灵**冬奥会**开幕式将激情上演
Text	新华社都灵2月9日电(记者丁莹阎涛)第20届**冬季奥运会**的开幕式将于当地时间10日晚8点在都灵奥林匹克体育场正式揭开神秘的面纱。

Indirect Associations in News Articles

Title	都灵冬奥会开幕式将激情上演
Text	新华社都灵2月9日电(记者丁莹阎涛)第20届冬季奥运会的开幕式将于当地时间10日晚8点在都灵奥林匹克体育场正式揭开神秘的面纱。

Indirect Associations in News Articles

Title 都灵冬奥会开幕式将激情上演

Text 新华社都灵2月9日电(记者丁莹阎涛)第20届冬季奥运会的开幕式将于当地时间10日晚8点在都灵奥林匹克体育场正式揭开神秘的面纱。

Indirect Associations in News Articles

Title	都灵冬奥会开幕式将激情上演
Text	新华社都灵2月9日电(记者丁莹阎涛)第20届冬季奥运会的开幕式将于当地时间10日晚8点在都灵奥林匹克体育场正式揭开神秘的面纱。

Indirect Associations in News Articles

Title	都灵冬奥会开幕式将激情上演
Text	新华社都灵2月9日电(记者丁莹阎涛)第20届冬季奥运会的开幕式将于当地时间10日晚8点在都灵奥林匹克体育场正式揭开神秘的面纱。

(冬季奥运会，冬奥会)

Winter Olympic Games

Mining Relations from the Web

Mining Relations from the Web

- **Identifying informal phrases**

Mining Relations from the Web

- **Identifying informal phrases**
- **For each informal phrase, identify its formal equivalent:**

Mining Relations from the Web

- **Identifying informal phrases**
- **For each informal phrase, identify its formal equivalent:**
  - Step-1: retrieving data from the web

Mining Relations from the Web

- **Identifying informal phrases**
- **For each informal phrase, identify its formal equivalent:**
  - Step-1: retrieving data from the web
  - Step-2: generating candidate hypotheses (i.e., formal phrases)

Mining Relations from the Web

- **Identifying informal phrases**
- **For each informal phrase, identify its formal equivalent:**
  - Step-1: retrieving data from the web
  - Step-2: generating candidate hypotheses (i.e., formal phrases)
  - Step-3: ranking hypotheses

Identifying informal phrases: bootstrapping

Identifying informal phrases: bootstrapping

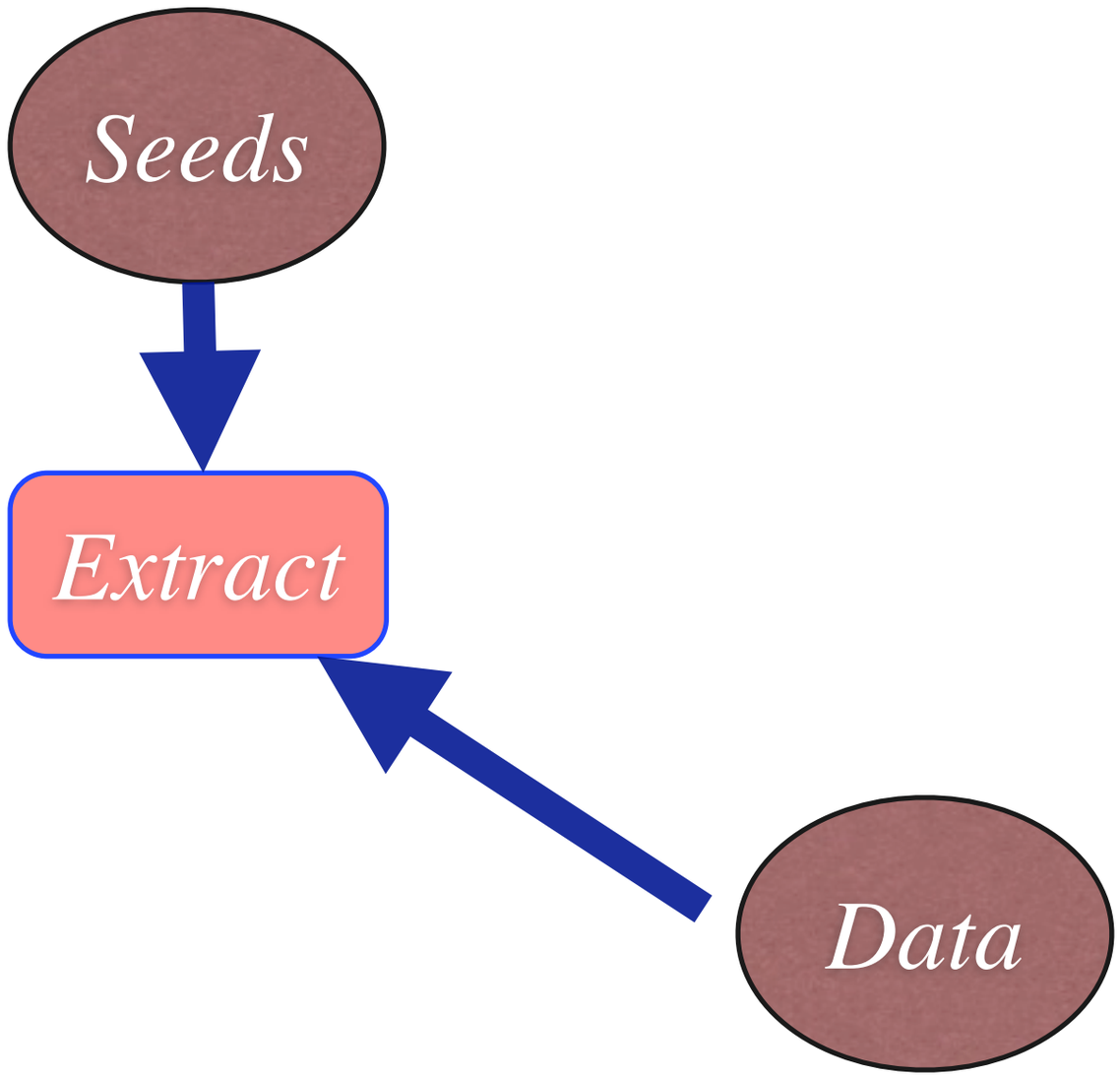
Seeds

Identifying informal phrases: bootstrapping

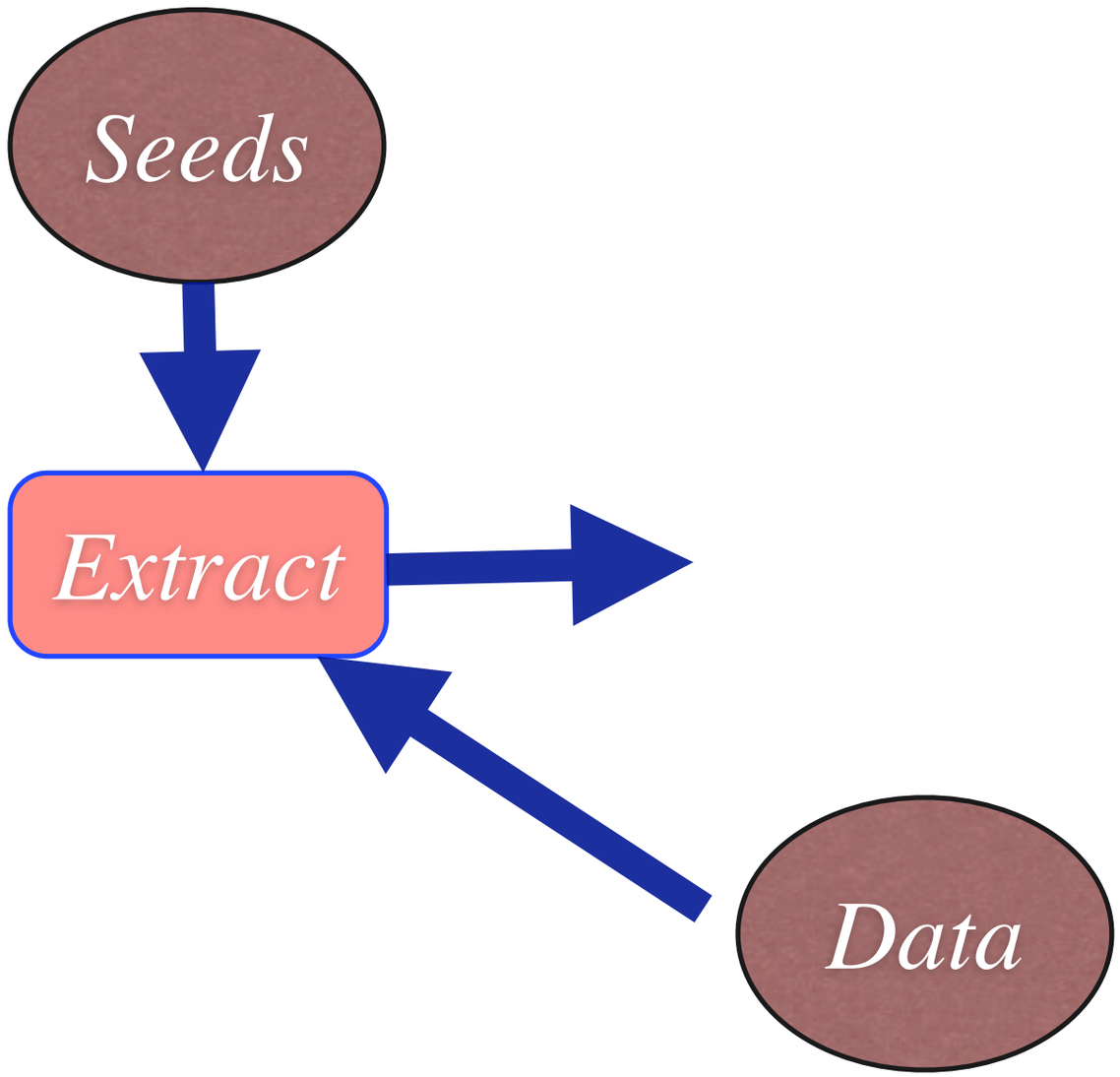
Seeds

Data

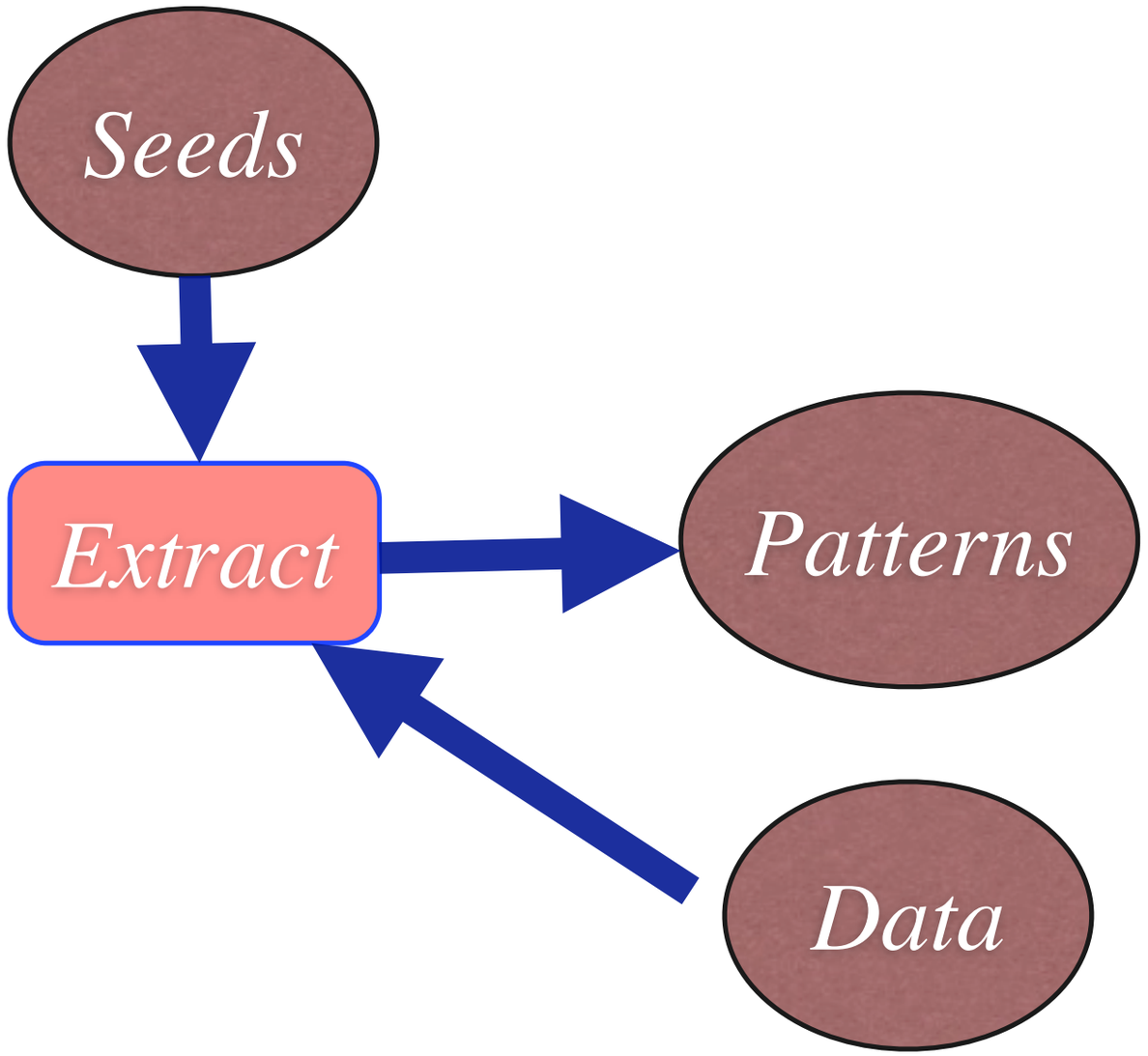
Identifying informal phrases: bootstrapping



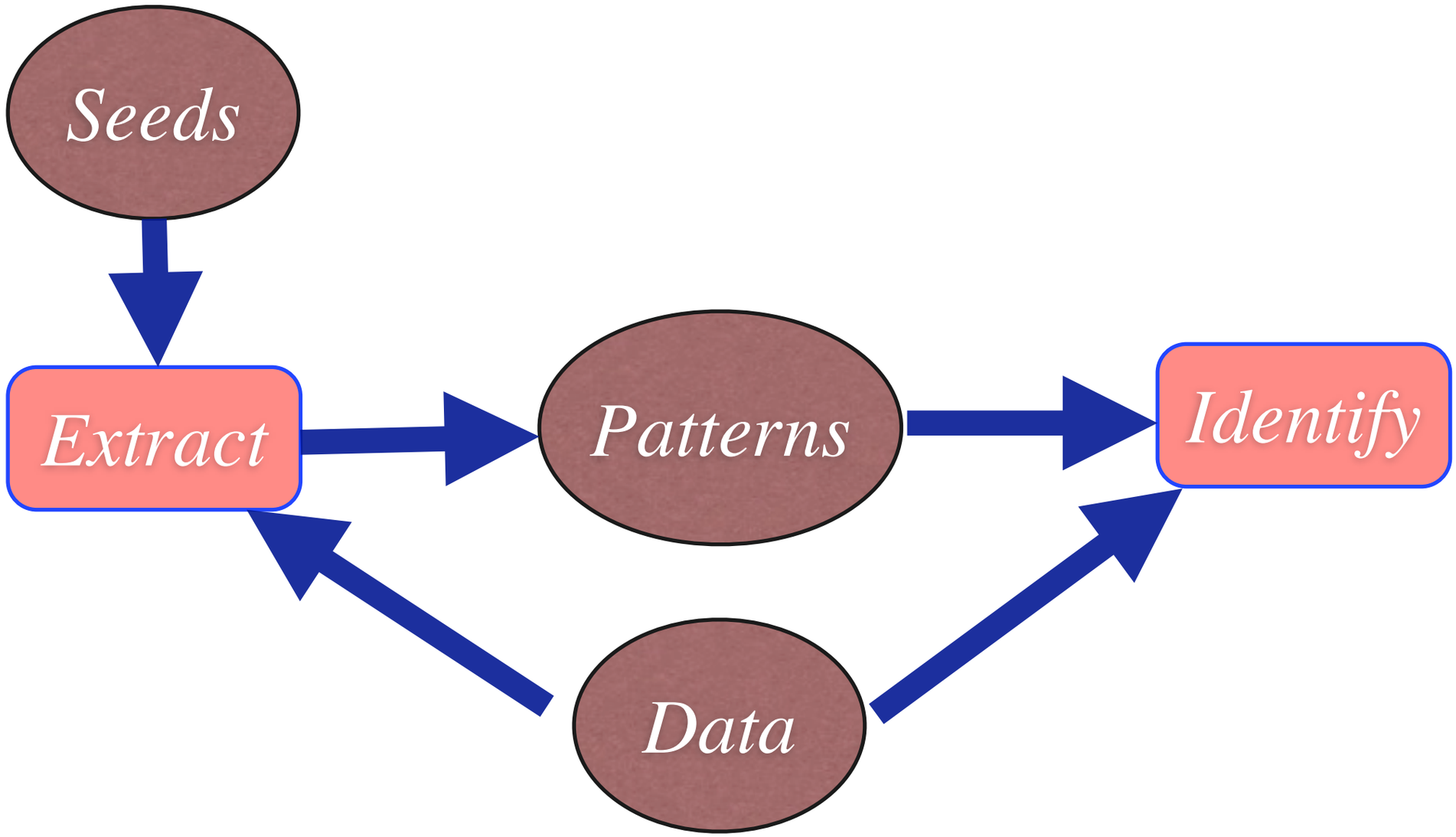
Identifying informal phrases: bootstrapping



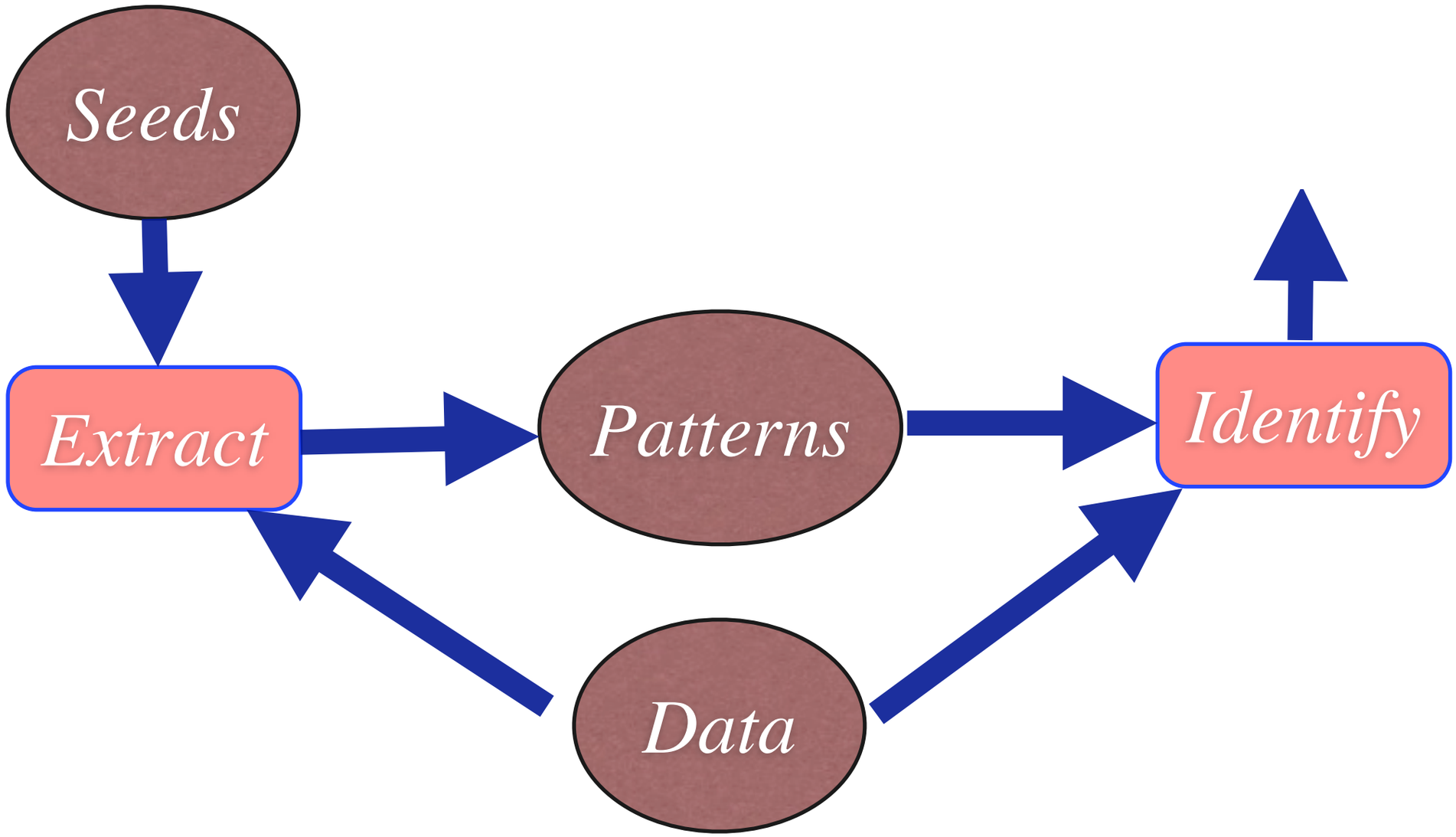
Identifying informal phrases: bootstrapping



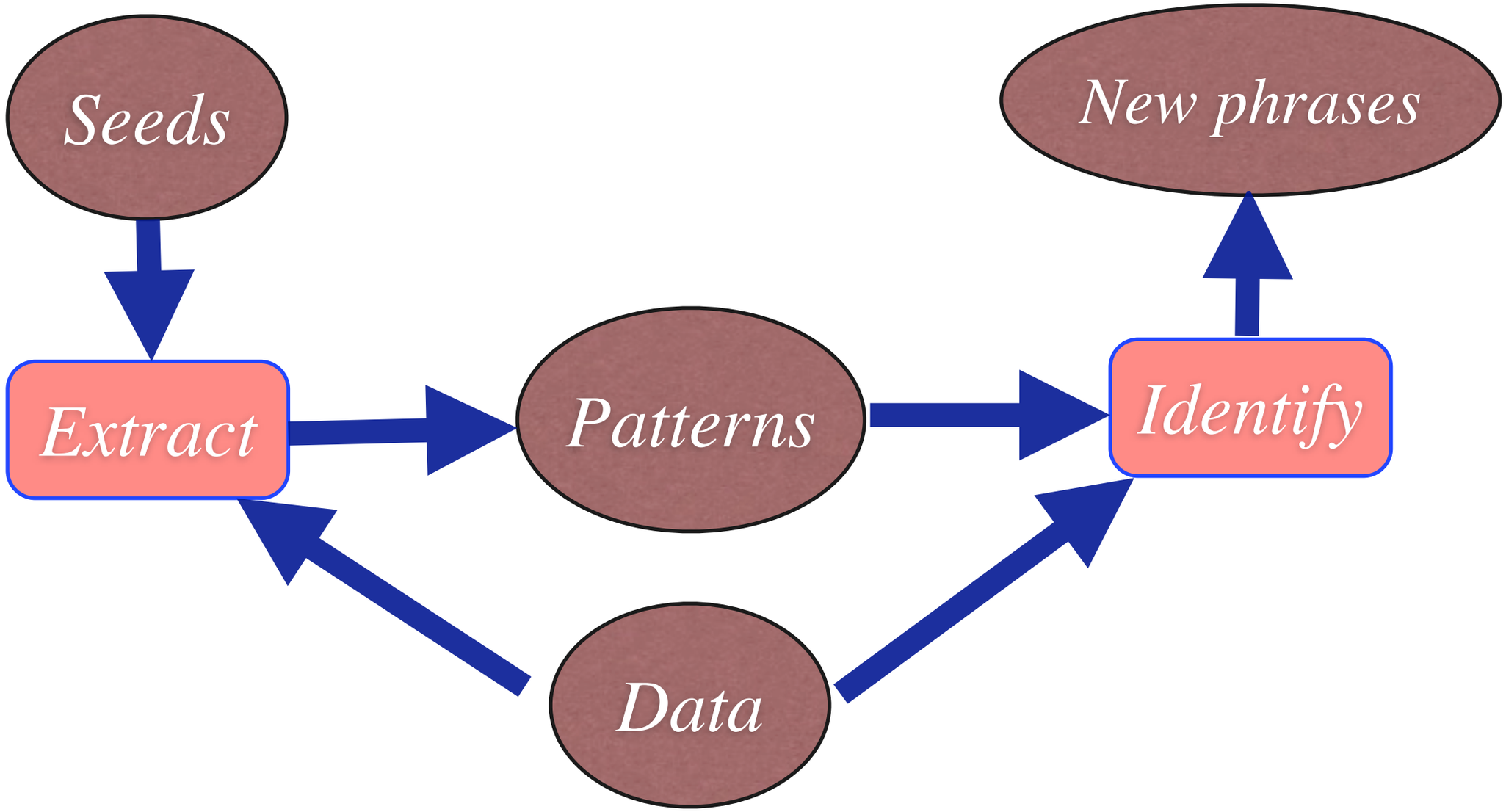
Identifying informal phrases: bootstrapping



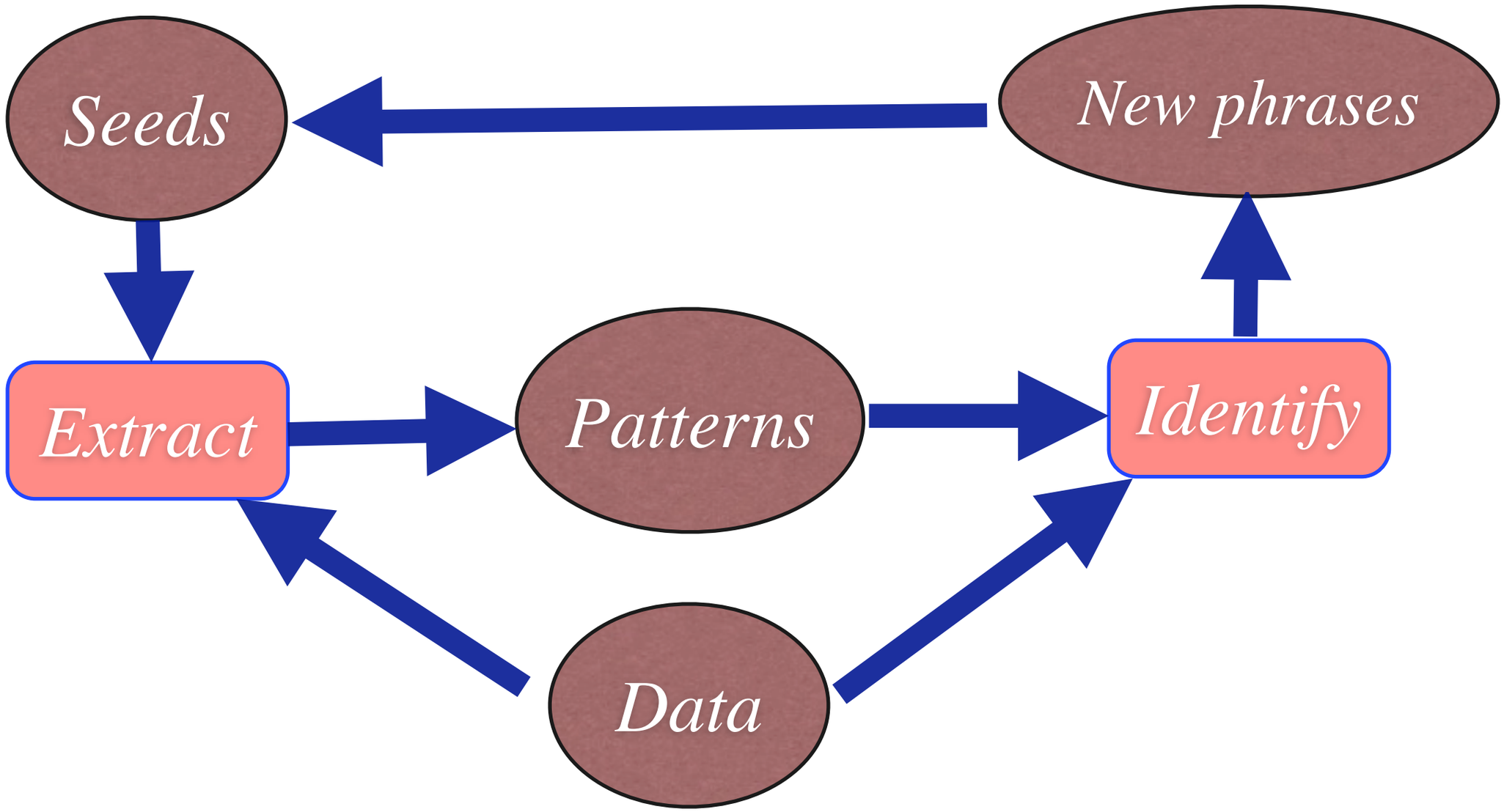
Identifying informal phrases: bootstrapping



Identifying informal phrases: bootstrapping



Identifying informal phrases: bootstrapping



Seed: (GF,女朋友)

Seed: (GF,女朋友)

*GF*是女朋友的意思

稀饭是喜欢的意思

这是他们的意思

Seed: (GF,女朋友)

GF是女朋友的意思

稀饭是喜欢的意思

这是他们的意思

Seed: (GF,女朋友)

GF是女朋友的意思

稀饭是喜欢的意思

这是他们的意思

Pattern:

*informal*是*formal*的意思

Seed: (GF,女朋友)

GF是女朋友的意思

稀饭是喜欢的意思

这是他们的意思

Pattern:

*informal*是*formal*的意思

New relations:

Seed: (GF,女朋友)

GF是女朋友的意思

稀饭是喜欢的意思

这是他们的意思

Pattern:

*informal*是*formal*的意思

New relations:

(稀饭, 喜欢)

(gruel, like)

Seed: (GF,女朋友)

GF是女朋友的意思

稀饭是喜欢的意思

这是他们的意思

Pattern:

*informal*是*formal*的意思

New relations:

(稀饭, 喜欢)　　(gruel, like)

(这, 他们)　　(this, they)

Seed: (GF,女朋友)

GF是女朋友的意思

稀饭是喜欢的意思

这是他们的意思

Pattern:

*informal*是*formal*的意思

New relations:

(稀饭, 喜欢)

(这, 他们)

(gruel, like)

(this, they)

Step-1: Retrieving data from the Web

- **Given an informal phrase, we retrieve data from the Web**
  - search the informal phrase using a search engine (e.g., *www.baidu.com*)
    - *just search the informal phrase*
    - *search the informal phrase with domain information*
    - *use a search engine dedicated to informal text, e.g., blogsearch.baidu.com*
  - download the relevant webpages returned by the search engine

新闻 网页 贴吧 知道 MP3 图片 视频

Bai du 百度

88

百度一下 结果中找 帮助 | 高级

把百度设为主页

u88连锁加盟网|u88致富有招|u88王刚代言|u88加盟连锁网|u88致富经..

国庆推出1.88万店型 藏源蒸疗 轻松蒸钱 我投资 你赚钱 中科院的祛痘特效药 农产品加工遍地黄金 好项目 致富只需860元 免收一切合作费用 国家扶持企业 上门建厂 成功付费 天地粮人鲜米坊 永和世家 彩色豆腐 0风险投资稳当赚钱 0风险...

www.u88.cn/ 125K 2008-10-25 - 百度快照

www.invest88.com 投资88网 |视频专家分析|音频专家建议|金融投资..

提供大陆及香港证券市场投资信息,提供在线音频证券投资建议及文字性的投资理财分析建议文章;并有部分香港股票,部分外汇产品及世界股票指数、贵金属的报价页面。

www.invest88.com/ 93K 2008-10-25 - 百度快照

88链自助链--Seo程序--友情链接联盟 - 无需审核，自动提取标题、...

88链旨在为站长提升更多更好的流量,对优秀网站免费加色、推荐。对作弊网站坚决屏蔽加入黑名单,永不收录;如果对屏蔽网站有异议可以加客服QQ申诉。本站拒 绝违犯中国法律法规的网站加入,谢谢合作! 自动加入说明 自动链更新日志 88链自助链--...

www.88link.cn/ 100K 2008-10-26 - 百度快照

叮当动漫-全国最大的免费卡通动漫网

钢之炼金术师[88] PSYREN[42] 某科学的超电磁炮[17] 番狗ナンバー 动漫音乐 劲爆 放送10周年记念ED 机动新选组-萌动之剑 全金属狂潮TSR 经典回忆灌篮高手音乐集 ひぐらしのなく顷にOP 凉宫ハルヒの忧郁ED 《最终幻想XII.和...

www.kt88.net/ 52K 2008-10-25 - 百度快照

求职 招聘 - Job88.com八方人才网

服务空间涉及:人才网站,传统招聘媒体,人才市场,猎头服务及跨境人才交流。

新闻 网页 贴吧 知道 MP3 图片 视频

Baidu 百度

88 网络用语

百度一下 结果中找 帮助 | 高级搜索

把百度设为主页

网络用语88，什么意思？ 百度知道

拜拜与88是谐音 回答者:蓝海梦露 - 魔法师 四级 2-3 20:07 网络流行语,或则网络用语比如88拜拜,现在又有些什... 哪些网络用语是"88"的意思? 请解释一下当前最流行的几个网络用语in high BT ... 收集网络聊天用语```知道的...

zhidao.baidu.com/question/3051118.html 12K 2007-5-6 - 百度快照

zhidao.baidu.com 上的更多结果

济宁一中88级2班，网络用语，QQ闪图，搞笑QQ表情，在线制作生成

网络用语 > 专用章的制作列表？多种不同风格的字体 所见所得的自由设计 支持阴影边框,颜色字体字大小 文字动画...QQ闪图下载_济宁一中88级2班 本页地址:http://www.zzxiu.com/html/2008/253/qqface_5331.shtml 论坛/博客贴图...

www.zzxiu.com/html/2008/253/qqface_5331.shtml 8K 2008-9-23 - 百度快照

网络用语大全 随意 My life

网络用语大全2007年04月17日 星期二 上午 09:50 PK就是单挑 粉丝 就是FANS 追星族 网络用语大全 看不懂不叫看不...886,88:再见 847:别生气 987:就不去,就不去 55555:哭 XXX:儿童不宜的东西 blah-blah:反复说 厚厚,吼吼,咔咔...

hi.baidu.com/4136697/blog/item/3d680f4c47 ... 26K 2007-4-20 - 百度快照

有趣的网络用语 爱问知识人

有趣的网络用语有趣的网络用语创建者:幻梦逍遥 创建时间:2007-02-18 01:44:19[3次点击] 如88用来指拜拜,1314有网络中用来指一生一世,而且8常用来代替不,比如8要即为不要,8给即为不给 类似的如酱紫即为这样子 还有一些比较常用...

iask.sina.com.cn/cidian/browse.php?name= ... 16K 2007-10-19 - 百度快照

百度 李娜吧 讲点百度贴吧上最常用的网络用语，免得一些同胞犯糊...

新闻 网页 贴吧 知道 MP3 图片 视频

Baidu 百度

88 网络用语

百度一下 结果中找 帮助 | 高级搜索

把百度设为主页

网络用语88，什么意思？ 百度知道

拜拜与88是谐音 回答者:蓝海梦露 - 魔法师 四级 2-3 20:07 网络流行语,或则网络用语比如88拜拜,现在又有些什... 哪些网络用语是"88"的意思? 请解释一下当前最流行的几个网络用语in high BT ... 收集网络聊天用语```知道的...

zhidao.baidu.com/question/3051118.html 12K 2007-5-6 - 百度快照

zhidao.baidu.com 上的更多结果

济宁一中88级2班，网络用语，QQ闪图，搞笑QQ表情，在线制作生成

网络用语 > 专用章的制作列表？多种不同风格的字体 所见所得的自由设计 支持阴影边框,颜色字体字大小 文字动画...QQ闪图下载_济宁一中88级2班 本页地址:http://www.zzxiu.com/html/2008/253/qqface_5331.shtml 论坛/博客贴图...

www.zzxiu.com/html/2008/253/qqface_5331.shtml 8K 2008-9-23 - 百度快照

网络用语大全 随意 My life

网络用语大全2007年04月17日 星期二 上午 09:50 PK就是单挑 粉丝 就是FANS 追星族 网络用语大全 看不懂不叫看不...886,88:再见 847:别生气 987:就不去,就不去 55555:哭 XXX:儿童不宜的东西 blah-blah:反复说 厚厚,吼吼,咔咔...

hi.baidu.com/4136697/blog/item/3d680f4c47 ... 26K 2007-4-20 - 百度快照

有趣的网络用语 爱问知识人

有趣的网络用语有趣的网络用语创建者:幻梦逍遥 创建时间:2007-02-18 01:44:19[3次点击] 如88用来指拜拜,1314有网络中用来指一生一世,而且8常用来代替不,比如8要即为不要,8给即为不给 类似的如酱紫即为这样子 还有一些比较常用...

iask.sina.com.cn/cidian/browse.php?name= ... 16K 2007-10-19 - 百度快照

百度 李娜吧 讲点百度贴吧上最常用的网络用语，免得一些同胞犯糊...

新闻 网页 贴吧 知道 MP3 图片 视频

Baidu 百度

88 网络用语

百度一下 结果中找 帮助 | 高级搜索

把百度设为主页

网络用语88，什么意思？ 百度知道

拜拜与88是谐音 回答者:蓝海梦露 - 魔法师 四级 2-3 20:07 网络流行语,或则网络用语比如88拜拜,现在又有些什... 哪些网络用语是"88"的意思? 请解释一下当前最流行的几个网络用语in high BT ... 收集网络聊天用语```知道的...

zhidao.baidu.com/question/3051118.html 12K 2007-5-6 - 百度快照

zhidao.baidu.com 上的更多结果

济宁一中88级2班，网络用语，QQ闪图，搞笑QQ表情，在线制作生成

网络用语 > 专用章的制作列表？多种不同风格的字体 所见所得的自由设计 支持阴影边框,颜色字体字大小 文字动画...QQ闪图下载_济宁一中88级2班 本页地址:http://www.zzxiu.com/html/2008/253/qqface_5331.shtml 论坛/博客贴图...

www.zzxiu.com/html/2008/253/qqface_5331.shtml 8K 2008-9-23 - 百度快照

网络用语大全 随意 My life

网络用语大全2007年04月17日 星期二 上午 09:50 PK就是单挑 粉丝 就是FANS 追星族 网络用语大全 看不懂不叫看不...886,88:再见 847:别生气 987:就不去,就不去 55555:哭 XXX:儿童不宜的东西 blah-blah:反复说 厚厚,吼吼,咔咔...

hi.baidu.com/4136697/blog/item/3d680f4c47 ... 26K 2007-4-20 - 百度快照

有趣的网络用语 爱问知识人

有趣的网络用语有趣的网络用语创建者:幻梦逍遥 创建时间:2007-02-18 01:44:19[3次点击] 如88用来指拜拜,1314有网络中用来指一生一世,而且8常用来代替不,比如8要即为不要,8给即为不给 类似的如酱紫即为这样子 还有一些比较常用...

iask.sina.com.cn/cidian/browse.php?name= ... 16K 2007-10-19 - 百度快照

百度 李娜吧 讲点百度贴吧上最常用的网络用语，免得一些同胞犯糊...

Step-2: Generating Candidate Hypotheses

- **Data-driven hypothesis generation**
  - we extract n-grams occurring within a certain distance from the informal phrase

Step-2: Generating Candidate Hypotheses

- **Data-driven hypothesis generation**
  - we extract n-grams occurring within a certain distance from the informal phrase

*GF*是女朋友的意思

Step-2: Generating Candidate Hypotheses

- **Data-driven hypothesis generation**
  - we extract n-grams occurring within a certain distance from the informal phrase

*GF*是女朋友的意思

Step-2: Generating Candidate Hypotheses

- **Data-driven hypothesis generation**
  - we extract n-grams occurring within a certain distance from the informal phrase

*GF*是女朋友的意思

Step-2: Generating Candidate Hypotheses

- **Data-driven hypothesis generation**
  - we extract n-grams occurring within a certain distance from the informal phrase

*GF*是女朋友的意思

1-gram
是
女
朋
友
的
意
思

Step-2: Generating Candidate Hypotheses

- **Data-driven hypothesis generation**
  - we extract n-grams occurring within a certain distance from the informal phrase

*GF*是女朋友的意思

1-gram	*2-gram*
是	是女
女	女朋
朋	朋友
友	友的
的	的意
意	意思
思	

Step-2: Generating Candidate Hypotheses

- **Data-driven hypothesis generation**
  - we extract n-grams occurring within a certain distance from the informal phrase

*GF*是女朋友的意思

1-gram	*2-gram*	*3-gram*
是	是女	是女朋
女	女朋	女朋友
朋	朋友	朋友的
友	友的	友的意
的	的意	的意思
意	意思	
思		

Step-2: Generating Candidate Hypotheses

- **Data-driven hypothesis generation**
  - we extract n-grams occurring within a certain distance from the informal phrase

*GF*是女朋友的意思

1-gram	*2-gram*	*3-gram*
是	是女	是女朋
女	女朋	女朋友
朋	朋友	朋友的
友	友的	友的意
的	的意	的意思
意	意思	
思		

Step-3: Ranking Hypotheses: Log-linear Models

- **Score of a hypothesis**

$$s(x,y) = \sum_{i=1}^{K} \Phi_i(x,y) \times \alpha_i$$

Step-3: Ranking Hypotheses: Log-linear Models

- **Score of a hypothesis**

$$s(x,y) = \sum_{i=1}^{K} \Phi_i(x,y) \times \alpha_i$$

- **Probability of a hypothesis**

$$P_{\vec{\alpha}}(y|x) = \frac{1}{Z(x,\vec{\alpha})} e^{s(x,y)}$$

Step-3: Ranking Hypotheses: Log-linear Models

- **Score of a hypothesis**

$$s(x,y) = \sum_{i=1}^{K} \Phi_i(x,y) \times \alpha_i$$

- **Probability of a hypothesis**

$$P_{\vec{\alpha}}(y|x) = \frac{1}{Z(x,\vec{\alpha})} e^{s(x,y)}$$

- **Regularized Conditional Log-likelihood**

$$LL_R(\vec{\alpha}) = \sum_{j=1}^{N} \log P_{\vec{\alpha}}(y_j|x_j) - \frac{||\vec{\alpha}||^2}{2\sigma^2}$$

Step-3: Ranking Hypotheses: Log-linear Models

- **Score of a hypothesis**

$$s(x,y) = \sum_{i=1}^{K} \Phi_i(x,y) \times \alpha_i$$

- **Probability of a hypothesis**

$$P_{\vec{\alpha}}(y|x) = \frac{1}{Z(x,\vec{\alpha})} e^{s(x,y)}$$

- **Regularized Conditional Log-likelihood**

$$LL_R(\vec{\alpha}) = \sum_{j=1}^{N} \log P_{\vec{\alpha}}(y_j|x_j) - \frac{||\vec{\alpha}||^2}{2\sigma^2}$$

- **Training**

$$\vec{\alpha}^* = \arg\max_{\vec{\alpha}} LL_R(\vec{\alpha})$$

Rule-driven Feature Functions

- **LD-PinYin(*x*,*y*)**
  - Levenshtein distance between PinYin of *x* and *y*
- **LEN-D-PinYin(*x*,*y*)**
  - Difference in the number of PinYin characters between *x* and *y*
- **Is-PinYin-Acronym(*x*,*y*)**
  - is *x* a PinYin acronym of *y*?
- **Is-CN-Abbreviation(*x*,*y*)**
  - is *x* a Chinese abbreviation of *y*?

Rule-driven Feature Functions

- **LD-PinYin(*x*,*y*)**
  - Levenshtein distance between PinYin of *x* and *y*
- **LEN-D-PinYin(*x*,*y*)**
  - Difference in the number of PinYin characters between *x* and *y*
- **Is-PinYin-Acronym(*x*,*y*)**
  - is *x* a PinYin acronym of *y*?
- **Is-CN-Abbreviation(*x*,*y*)**
  - is *x* a Chinese abbreviation of *y*?

Is-PinYin-Acronym(GG, 哥哥)=1,
Is-PinYin-Acronym(GG, 兄弟)=0.

Rule-driven Feature Functions

- **LD-PinYin(*x*,*y*)**
  - Levenshtein distance between PinYin of *x* and *y*
- **LEN-D-PinYin(*x*,*y*)**
  - Difference in the number of PinYin characters between *x* and *y*
- **Is-PinYin-Acronym(*x*,*y*)**
  - is *x* a PinYin acronym of *y*?

Is-PinYin-Acronym(GG, 哥哥)=1,
Is-PinYin-Acronym(GG, 兄弟)=0.

- **Is-CN-Abbreviation(*x*,*y*)**
  - is *x* a Chinese abbreviation of *y*?

Is-CN-Abbreviation(美军,美国军队)=1,
Is-CN-Abbreviation(美军,中国军队)=0.

Data-driven Feature Functions

- **n-gram co-occurrence relative frequency**
  - how often *x* and *y* appear together?
- **Feature on a definition pattern**
  - Does co-occurrence of *x* and *y* follow a definition pattern?
- **Feature on the number of relevant webpages**
  - search the web using the pair of *x* and *y*
  - get the number of webpages returned

Results on Bootstrapping Informal Phrases

Size of seed set	130
Size of candidate set	3000
Size of test set	750
Recall	**30%**

Precisions on Relation Extraction

Category		Precision (%)			
		Top-1	Top-10	Top-50	Top-100
Homophone	Same PinYin	63.2	73.7	84.2	84.2
	Similar PinYin	40.0	60.0	70.0	80.0
	Number	81.1	91.6	95.8	96.8
Abbreviation	Chinese abbreviation	11.8	41.2	52.9	52.9
Acronym	PinYin Acronym	82.1	94.6	96.4	96.4
	English Acronym	21.9	46.9	56.3	59.4
Transliteration		20.0	40.0	50.0	50.0
Average		**61.8**	**77.1**	**83.1**	**84.7**

Precisions on Relation Extraction

Category		Precision (%)			
		Top-1	Top-10	Top-50	Top-100
Homophone	Same PinYin	63.2	73.7	84.2	84.2
	Similar PinYin	40.0	60.0	70.0	80.0
	Number	81.1	91.6	95.8	96.8
Abbreviation	Chinese abbreviation	11.8	41.2	52.9	52.9
Acronym	PinYin Acronym	82.1	94.6	96.4	96.4
	English Acronym	21.9	46.9	56.3	59.4
Transliteration		20.0	40.0	50.0	50.0
Average		**61.8**	**77.1**	**83.1**	**84.7**

Precisions on Relation Extraction

Category		Precision (%)			
		Top-1	Top-10	Top-50	Top-100
Homophone	Same PinYin	63.2	73.7	84.2	84.2
	Similar PinYin	40.0	60.0	70.0	80.0
	Number	81.1	91.6	95.8	96.8
Abbreviation	Chinese abbreviation	11.8	41.2	52.9	52.9
Acronym	PinYin Acronym	82.1	94.6	96.4	96.4
	English Acronym	21.9	46.9	56.3	59.4
Transliteration		20.0	40.0	50.0	50.0
Average		**61.8**	**77.1**	**83.1**	**84.7**

Precisions on Relation Extraction

Category		Precision (%)			
		Top-1	Top-10	Top-50	Top-100
Homophone	Same PinYin	63.2	73.7	84.2	84.2
	Similar PinYin	40.0	60.0	70.0	80.0
	Number	81.1	91.6	95.8	96.8
Abbreviation	Chinese abbreviation	11.8	41.2	52.9	52.9
Acronym	PinYin Acronym	82.1	94.6	96.4	96.4
	English Acronym	21.9	46.9	56.3	59.4
Transliteration		20.0	40.0	50.0	50.0
Average		**61.8**	**77.1**	**83.1**	**84.7**

Average (using rule-driven features)	*26.1*	*53.4*	*66.3*	*72.3*

Precisions on Relation Extraction

Category		Precision (%)			
		Top-1	Top-10	Top-50	Top-100
Homophone	Same PinYin	63.2	73.7	84.2	84.2
	Similar PinYin	40.0	60.0	70.0	80.0
	Number	81.1	91.6	95.8	96.8
Abbreviation	Chinese abbreviation	11.8	41.2	52.9	52.9
Acronym	PinYin Acronym	82.1	94.6	96.4	96.4
	English Acronym	21.9	46.9	56.3	59.4
Transliteration		20.0	40.0	50.0	50.0
Average		**61.8**	**77.1**	**83.1**	**84.7**
Average (using rule-driven features)		*26.1*	*53.4*	*66.3*	*72.3*
Average (using data-driven features)		*51.4*	*71.1*	*81.1*	*82.7*

Summary

- We proposed an approach to extract relations between informal and formal Chinese phrases from Web corpora
- Our approach combines both rule and data-driven features
- The combinations of both rule and data-driven features outperforms the case just using only rule or data-driven features

Training and test examples are online at

http://www.cs.jhu.edu/~zfli

Thank you!