IJCAI 2015

Iterative Learning of Parallel Lexicons and Phrases from Non-Parallel Corpora

Meiping Dong, Yang Liu, Huanbo Luan, Maosong Sun,
Tatsuya Izuha, and Dakun Zhang

Learning Lexicons from Parallel Corpora

- Parallel corpora are important for learning lexicons

Garcia y asociados

Garcia and associates

sus asociados no son fuertes

his associates are not strong

Garcia y sus asociados no son enemigos

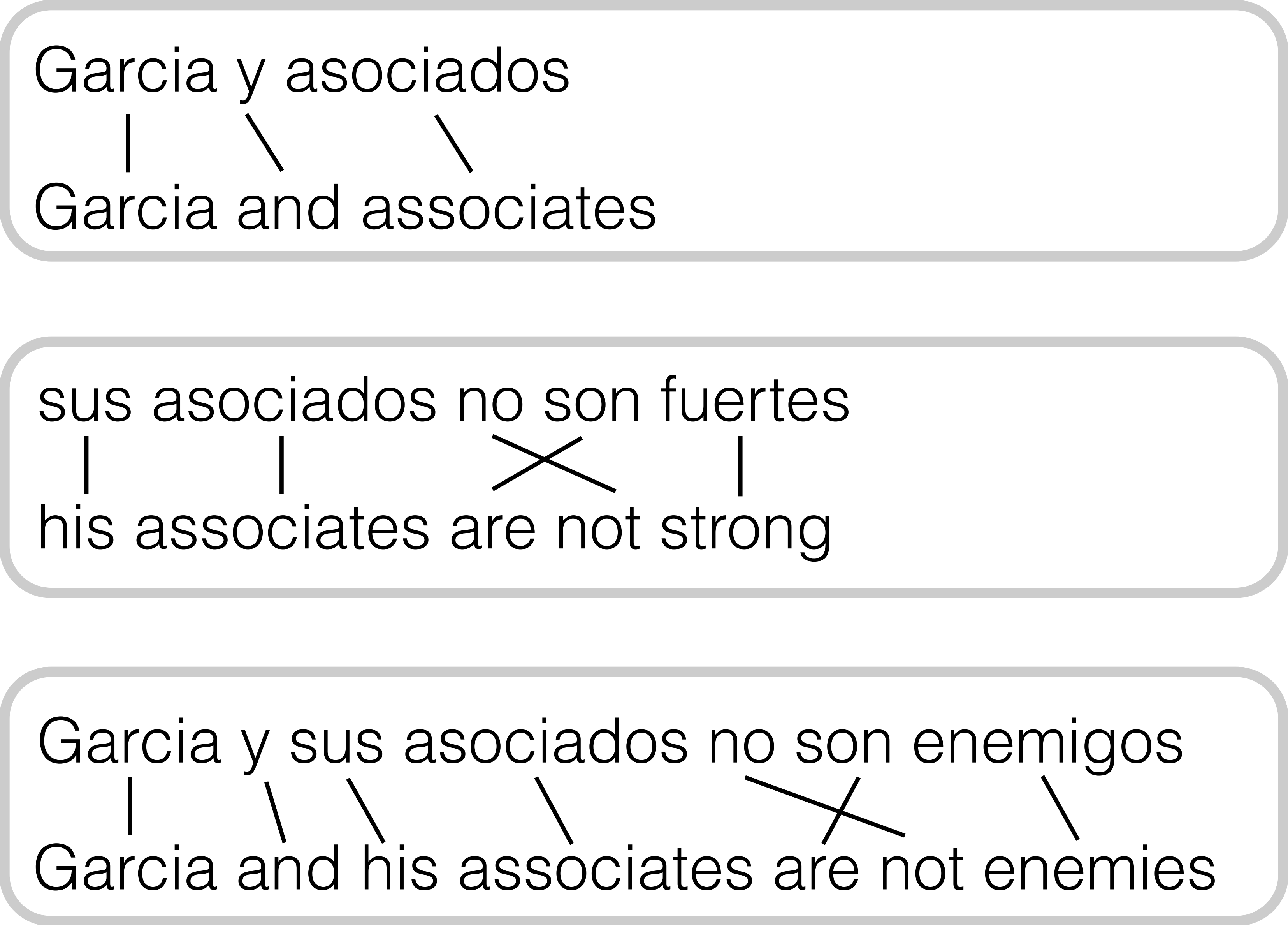
Garcia and his associates are not enemies

parallel corpora

(Brown et al., 1993)

Learning Lexicons from Parallel Corpora

- Parallel corpora are important for learning lexicons



Garcia y asociados

Garcia and associates

sus asociados no son fuertes

his associates are not strong

Garcia y sus asociados no son enemigos

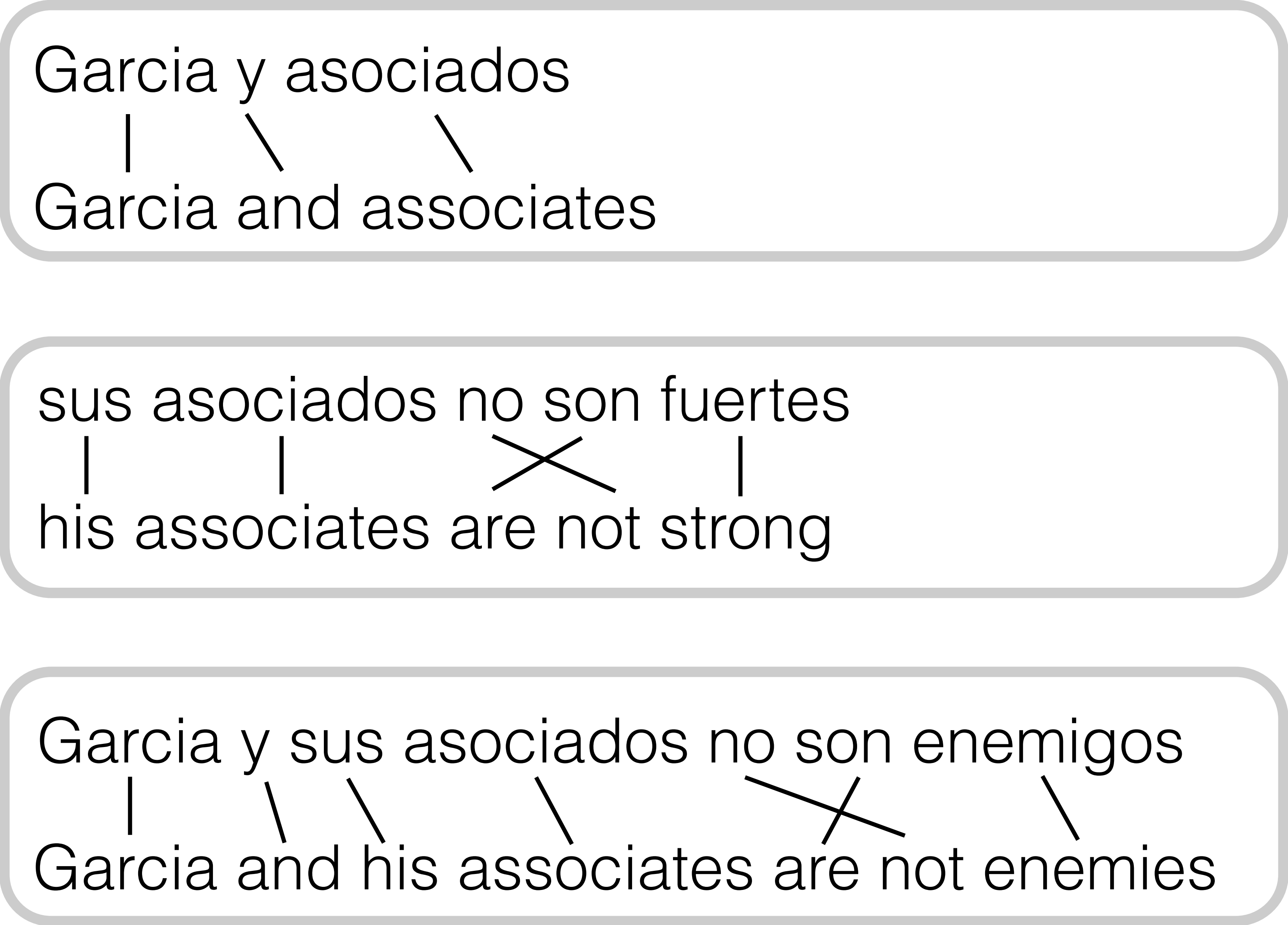
Garcia and his associates are not enemies

parallel corpora

(Brown et al., 1993)

Learning Lexicons from Parallel Corpora

- Parallel corpora are important for learning lexicons



Garcia y asociados

Garcia and associates

sus asociados no son fuertes

his associates are not strong

Garcia y sus asociados no son enemigos

Garcia and his associates are not enemies

parallel corpora

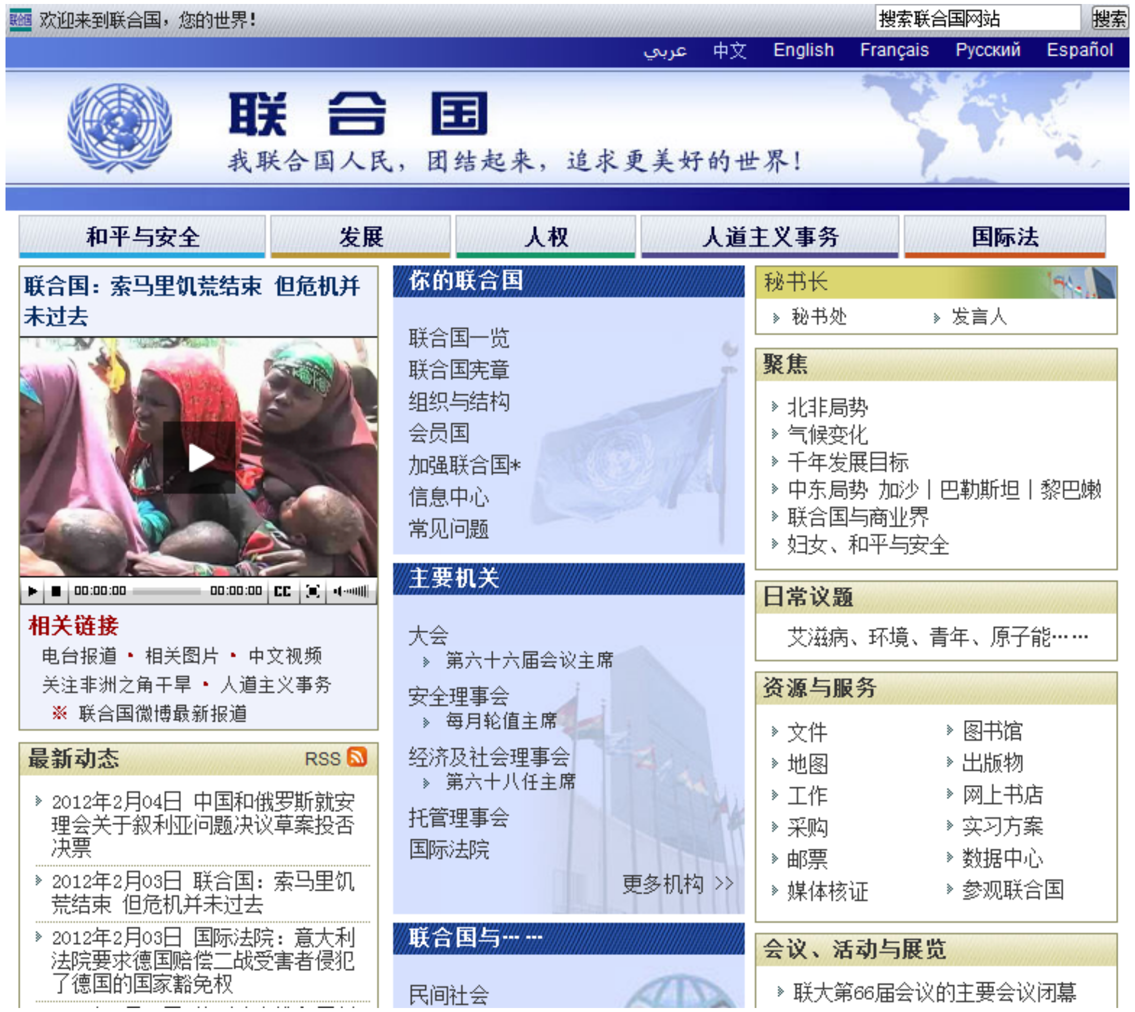
Spanish	English
Garcia	Garcia
y	and
asociados	associates
sus	his
no	not
son	are
fuertes	strong
enemigos	enemies

parallel lexicon

(Brown et al., 1993)

Limitation of Parallel Corpora

- Parallel corpora are limited in both quantity and coverage



欢迎来到联合国，您的世界！

搜索联合国网站 搜索

عربي 中文 English Français Русский Español

联 合 国

我联合国人民，团结起来，追求更美好的世界！

和平与安全 发展 人权 人道主义事务 国际法

联合国：索马里饥荒结束 但危机并未过去

▶ ■ 00:00:00 00:00:00 CC

相关链接

电台报道 • 相关图片 • 中文视频

关注非洲之角干旱 • 人道主义事务

※ 联合国微博最新报道

最新动态 RSS

- 2012年2月04日 中国和俄罗斯就安理会关于叙利亚问题决议草案投否决票
- 2012年2月03日 联合国：索马里饥荒结束 但危机并未过去
- 2012年2月03日 国际法院：意大利法院要求德国赔偿二战受害者侵犯了德国的国家豁免权

你的联合国

联合国一览
联合国宪章
组织与结构
会员国
加强联合国*
信息中心
常见问题

主要机关

大会
- 第六十六届会议主席

安全理事会
- 每月轮值主席

经济及社会理事会
- 第六十八任主席

托管理事会
国际法院

更多机构 >>

联合国与… …

民间社会

秘书长

- 秘书处
- 发言人

聚焦

- 北非局势
- 气候变化
- 千年发展目标
- 中东局势 加沙 | 巴勒斯坦 | 黎巴嫩
- 联合国与商业界
- 妇女、和平与安全

日常议题

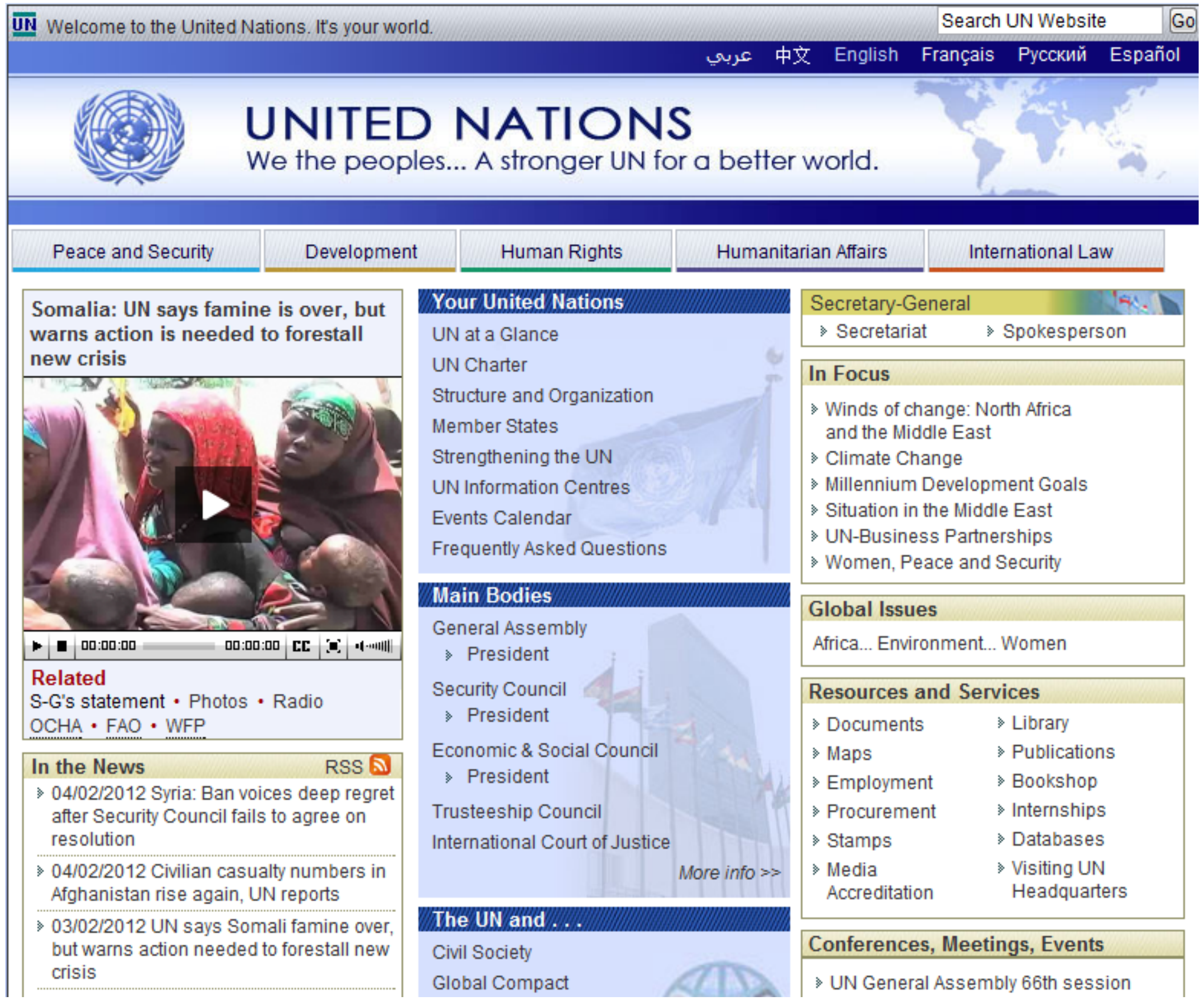
艾滋病、环境、青年、原子能……

资源与服务

文件	图书馆
地图	出版物
工作	网上书店
采购	实习方案
邮票	数据中心
媒体核证	参观联合国

会议、活动与展览

- 联大第66届会议的主要会议闭幕



Welcome to the United Nations. It's your world.

Search UN Website Go

عربي 中文 English Français Русский Español

UNITED NATIONS

We the peoples... A stronger UN for a better world.

Peace and Security Development Human Rights Humanitarian Affairs International Law

Somalia: UN says famine is over, but warns action is needed to forestall new crisis

▶ ■ 00:00:00 00:00:00 CC

Related

S-G's statement • Photos • Radio

OCHA • FAO • WFP

In the News RSS

- 04/02/2012 Syria: Ban voices deep regret after Security Council fails to agree on resolution
- 04/02/2012 Civilian casualty numbers in Afghanistan rise again, UN reports
- 03/02/2012 UN says Somali famine over, but warns action needed to forestall new crisis

Your United Nations

UN at a Glance
UN Charter
Structure and Organization
Member States
Strengthening the UN
UN Information Centres
Events Calendar
Frequently Asked Questions

Main Bodies

General Assembly
- President

Security Council
- President

Economic & Social Council
- President

Trusteeship Council
International Court of Justice

More info >>

The UN and . . .

Civil Society
Global Compact

Secretary-General

- Secretariat
- Spokesperson

In Focus

- Winds of change: North Africa and the Middle East
- Climate Change
- Millennium Development Goals
- Situation in the Middle East
- UN-Business Partnerships
- Women, Peace and Security

Global Issues

Africa... Environment... Women

Resources and Services

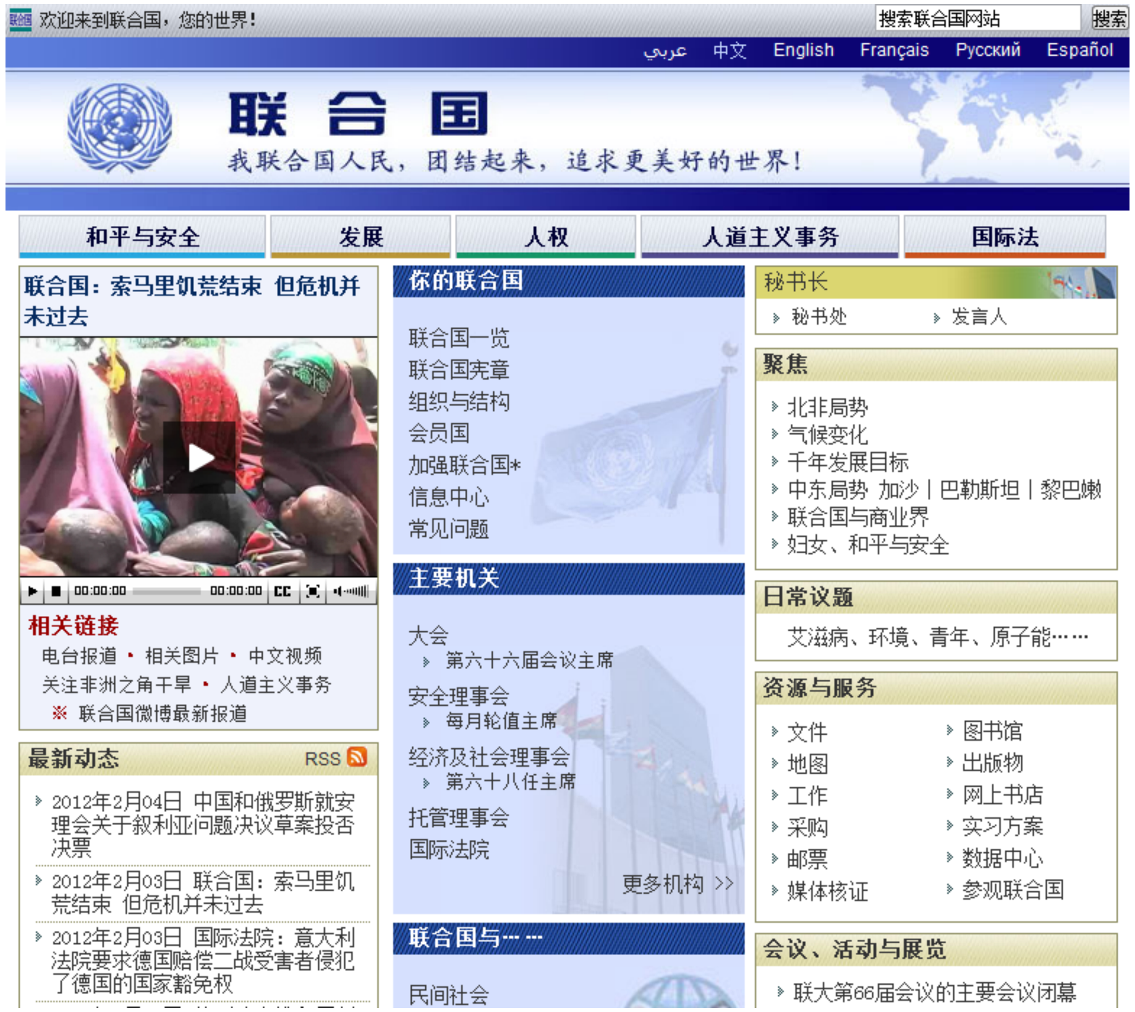
Documents	Library
Maps	Publications
Employment	Bookshop
Procurement	Internships
Stamps	Databases
Media Accreditation	Visiting UN Headquarters

Conferences, Meetings, Events

- UN General Assembly 66th session

Limitation of Parallel Corpora

- Parallel corpora are limited in both quantity and coverage



欢迎来到联合国，您的世界！ 搜索联合国网站 搜索

عربي 中文 English Français Русский Español

联 合 国

我联合国人民，团结起来，追求更美好的世界！

和平与安全 | 发展 | 人权 | 人道主义事务 | 国际法

联合国：索马里饥荒结束 但危机并未过去

00:00:00 00:00:00 CC

相关链接

电台报道 • 相关图片 • 中文视频

关注非洲之角干旱 • 人道主义事务

※ 联合国微博最新报道

最新动态 RSS

- 2012年2月04日 中国和俄罗斯就安理会关于叙利亚问题决议草案投否决票
- 2012年2月03日 联合国：索马里饥荒结束 但危机并未过去
- 2012年2月03日 国际法院：意大利法院要求德国赔偿二战受害者侵犯了德国的国家豁免权

你的联合国

联合国一览
联合国宪章
组织与结构
会员国
加强联合国*
信息中心
常见问题

主要机关

大会
- 第六十六届会议主席

安全理事会
- 每月轮值主席

经济及社会理事会
- 第六十八任主席

托管理事会
国际法院

更多机构 >>

联合国与--- ---

民间社会

秘书长

- 秘书处
- 发言人

聚焦

- 北非局势
- 气候变化
- 千年发展目标
- 中东局势 加沙 | 巴勒斯坦 | 黎巴嫩
- 联合国与商业界
- 妇女、和平与安全

日常议题

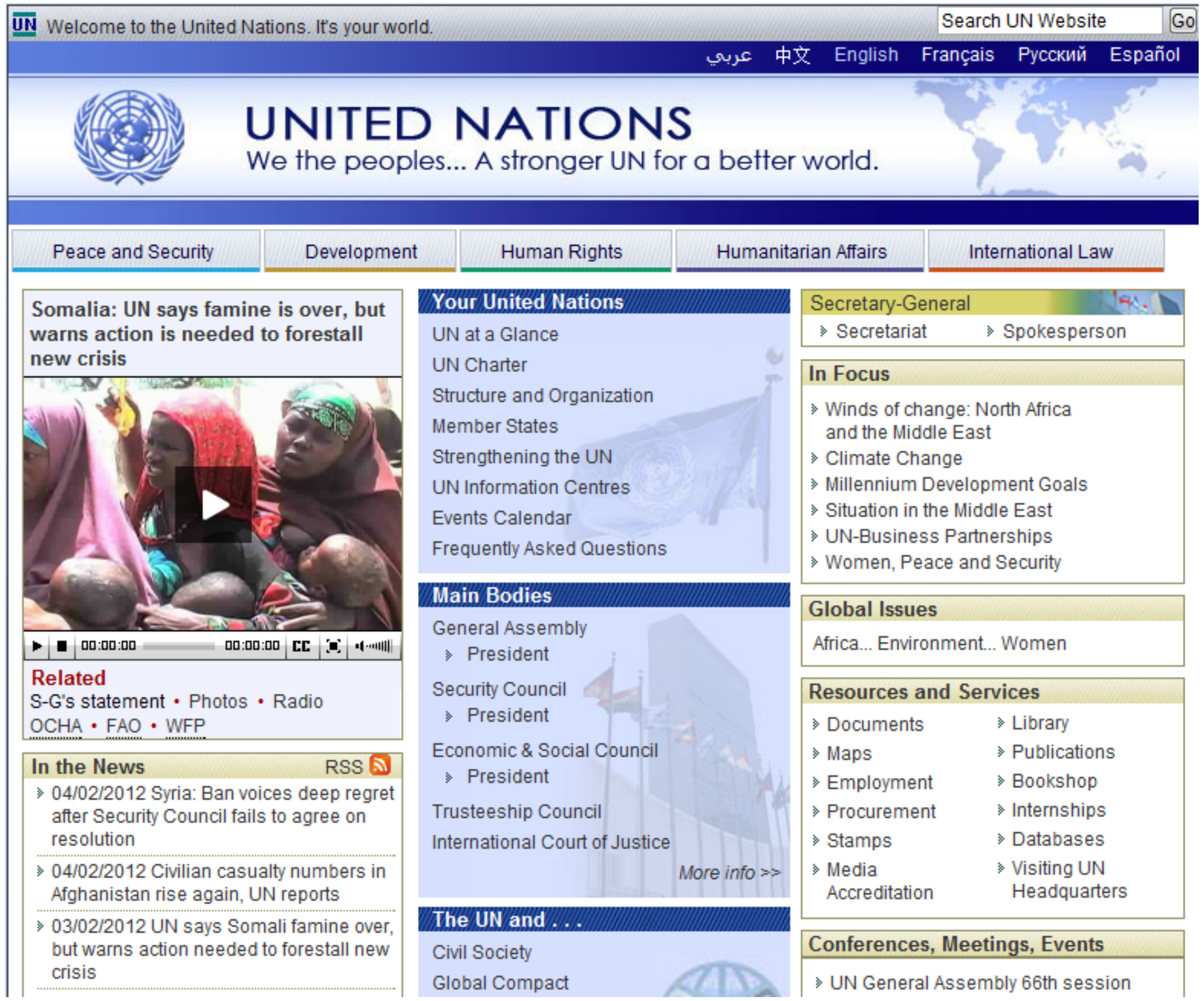
艾滋病、环境、青年、原子能……

资源与服务

- 文件
- 地图
- 工作
- 采购
- 邮票
- 媒体核证
- 图书馆
- 出版物
- 网上书店
- 实习方案
- 数据中心
- 参观联合国

会议、活动与展览

- 联大第66届会议的主要会议闭幕



UN Welcome to the United Nations. It's your world. Search UN Website Go

عربي 中文 English Français Русский Español

UNITED NATIONS

We the peoples... A stronger UN for a better world.

Peace and Security | Development | Human Rights | Humanitarian Affairs | International Law

Somalia: UN says famine is over, but warns action is needed to forestall new crisis

00:00:00 00:00:00 CC

Related

S-G's statement • Photos • Radio

OCHA • FAO • WFP

In the News RSS

- 04/02/2012 Syria: Ban voices deep regret after Security Council fails to agree on resolution
- 04/02/2012 Civilian casualty numbers in Afghanistan rise again, UN reports
- 03/02/2012 UN says Somali famine over, but warns action needed to forestall new crisis

Your United Nations

UN at a Glance
UN Charter
Structure and Organization
Member States
Strengthening the UN
UN Information Centres
Events Calendar
Frequently Asked Questions

Main Bodies

General Assembly
- President

Security Council
- President

Economic & Social Council
- President

Trusteeship Council
International Court of Justice

More info >>

The UN and . . .

Civil Society
Global Compact

Secretary-General

- Secretariat
- Spokesperson

In Focus

- Winds of change: North Africa and the Middle East
- Climate Change
- Millennium Development Goals
- Situation in the Middle East
- UN-Business Partnerships
- Women, Peace and Security

Global Issues

Africa... Environment... Women

Resources and Services

- Documents
- Maps
- Employment
- Procurement
- Stamps
- Media Accreditation
- Library
- Publications
- Bookshop
- Internships
- Databases
- Visiting UN Headquarters

Conferences, Meetings, Events

- UN General Assembly 66th session

Can we learn lexicons from non-parallel corpora?

Previous Work

- Joint parallel sentence and lexicon extraction (Fung and Cheung, 2004)
- Parallel sub-sentential fragment extraction (Munteanu and Marcu, 2006; Quirk et al., 2007; Cettolo et al., 2010)
- Extracting parallel phrases using lexicons (Zhang and Zong, 2013)
- Translation as decipherments (Ravi and Knight, 2011; Nuhn et al., 2012; Dou and Knight, 2012)
- Transductive learning on monolingual corpora (Ueffing et al., 2007; Bertoldi and Federico, 2009)

Our Work

- Learning lexicons and phrases from non-parallel corpora

Garcia y asociados

sus asociados no son fuertes

no son enemigos

Buenos Ares

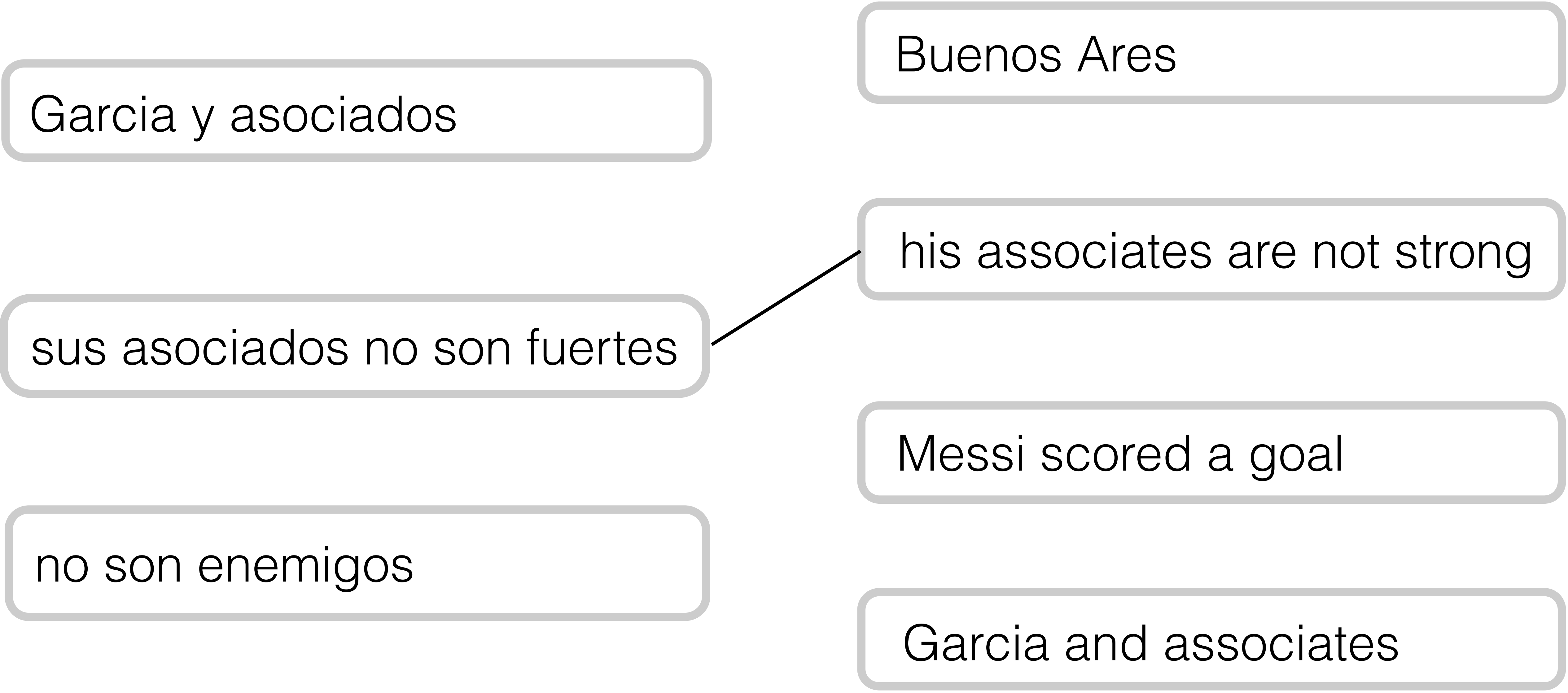
his associates are not strong

Messi scored a goal

Garcia and associates

Our Work

- Learning lexicons and phrases from non-parallel corpora



Garcia y asociados

sus asociados no son fuertes

no son enemigos

Buenos Ares

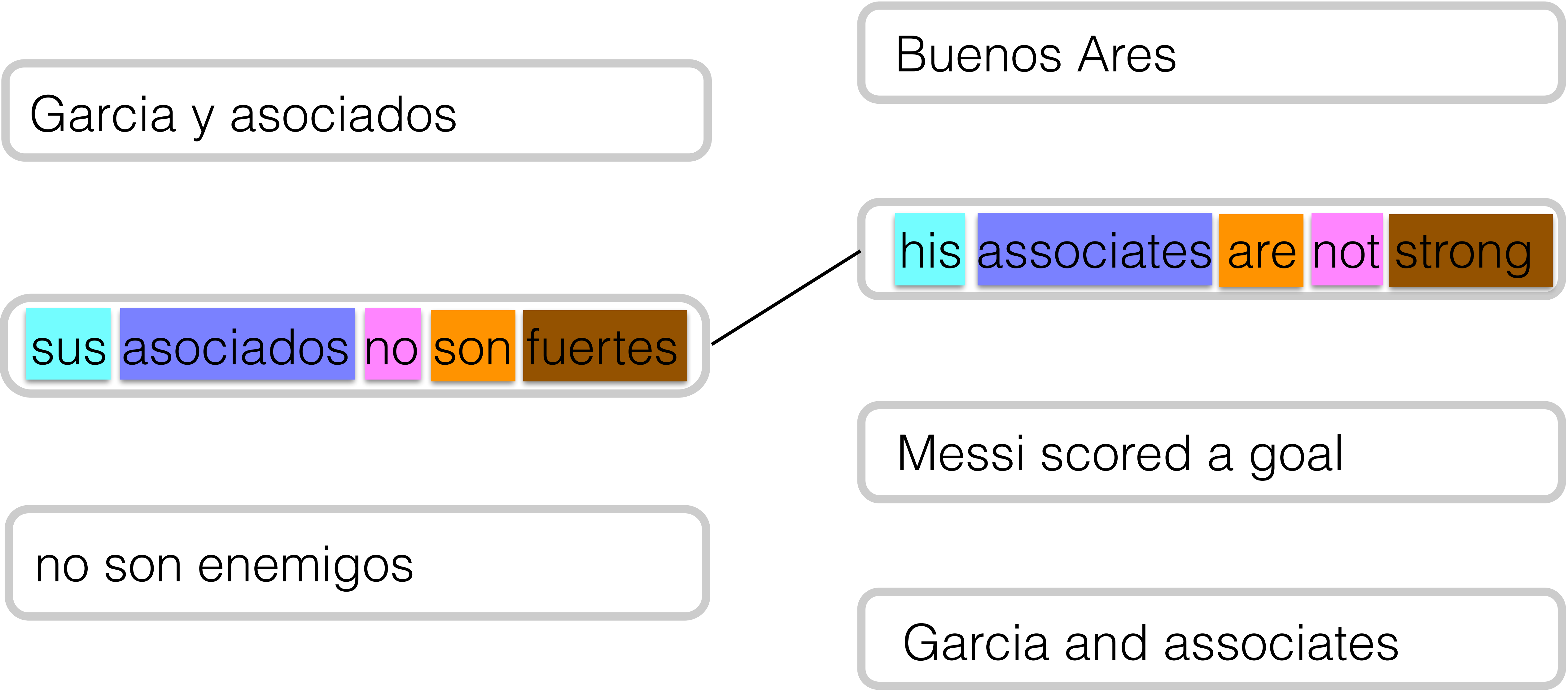
his associates are not strong

Messi scored a goal

Garcia and associates

Our Work

- Learning lexicons and phrases from non-parallel corpora



Garcia y asociados

sus asociados no son fuertes

no son enemigos

Buenos Ares

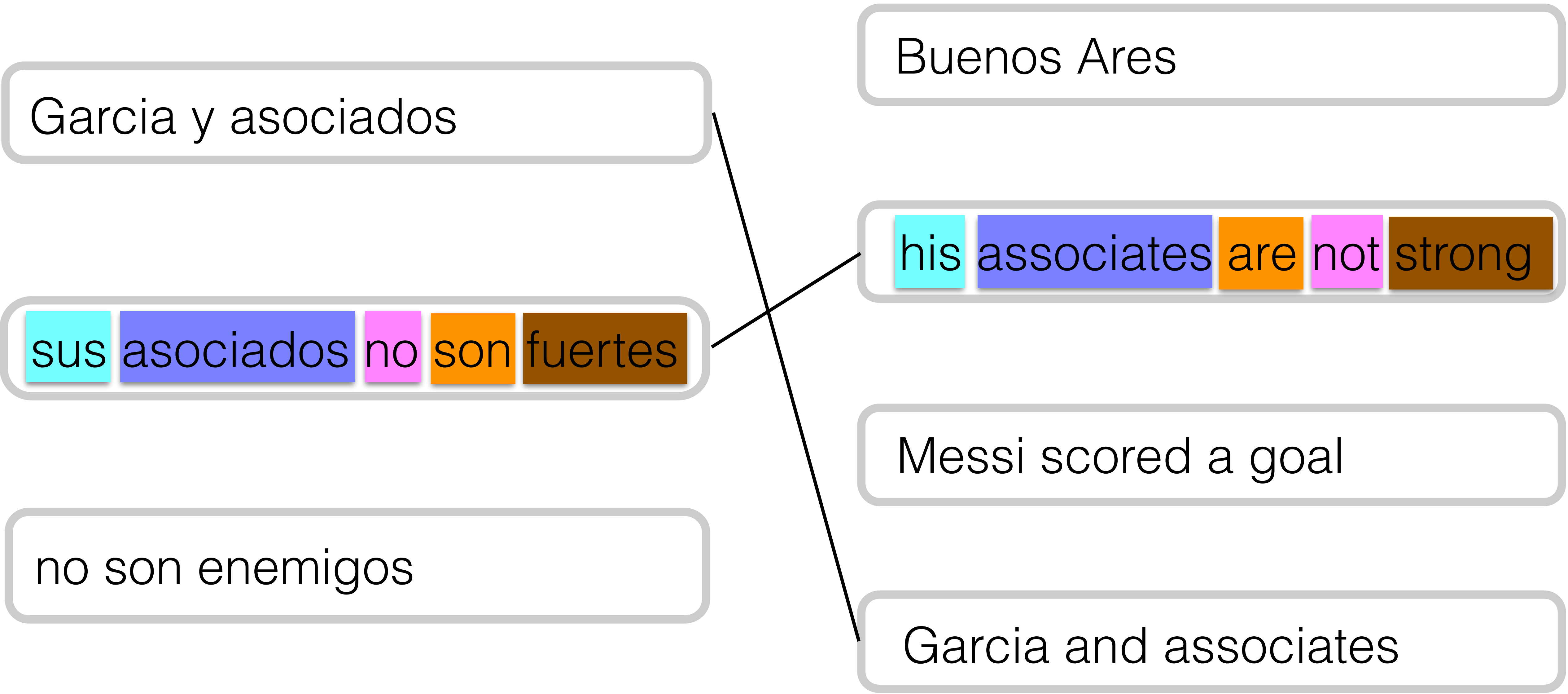
his associates are not strong

Messi scored a goal

Garcia and associates

Our Work

- Learning lexicons and phrases from non-parallel corpora



Garcia y asociados

sus asociados no son fuertes

no son enemigos

Buenos Ares

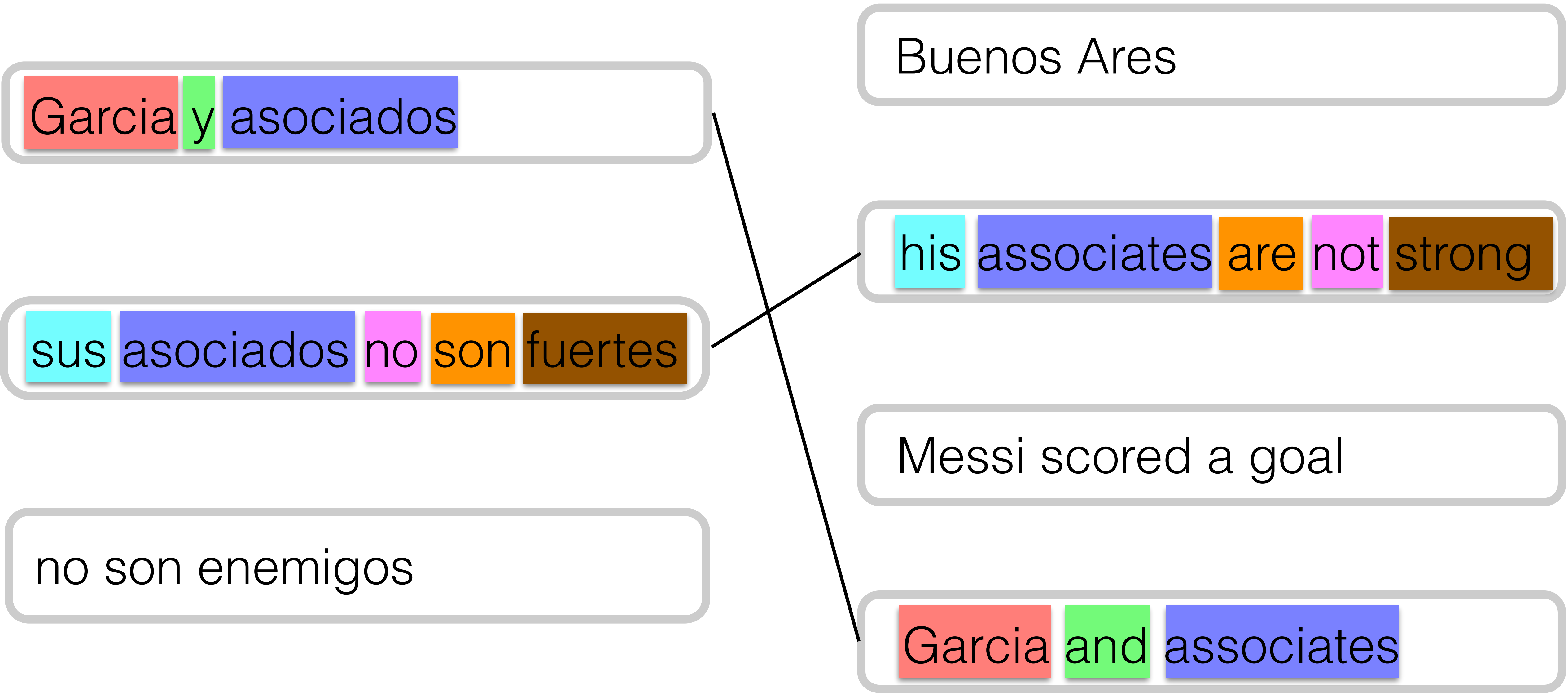
his associates are not strong

Messi scored a goal

Garcia and associates

Our Work

- Learning lexicons and phrases from non-parallel corpora



Garcia y asociados

sus asociados no son fuertes

no son enemigos

Buenos Ares

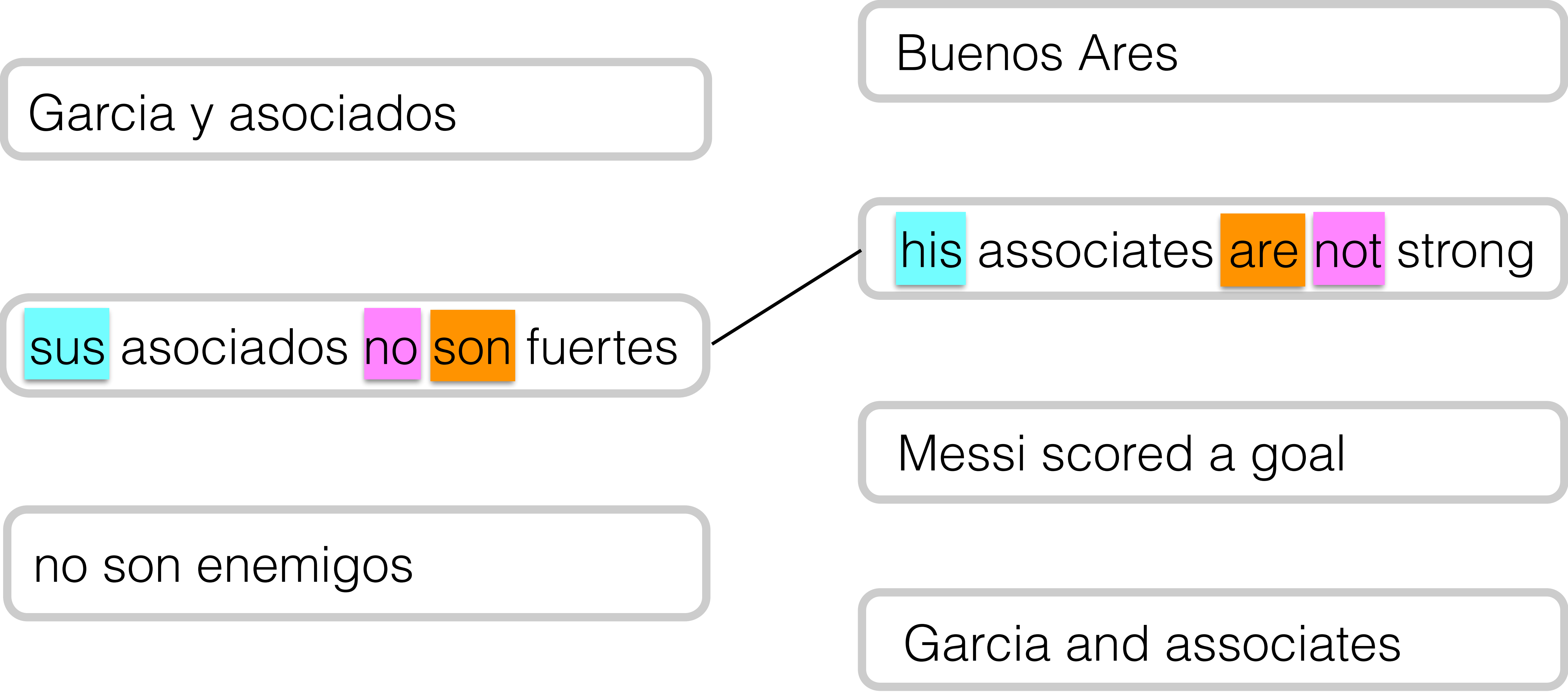
his associates are not strong

Messi scored a goal

Garcia and associates

Idea

- “**find one, get more**” (Fung and Cheung, 2004) **iteratively**



Buenos Ares

Garcia y asociados

his associates are not strong

sus asociados no son fuertes

Messi scored a goal

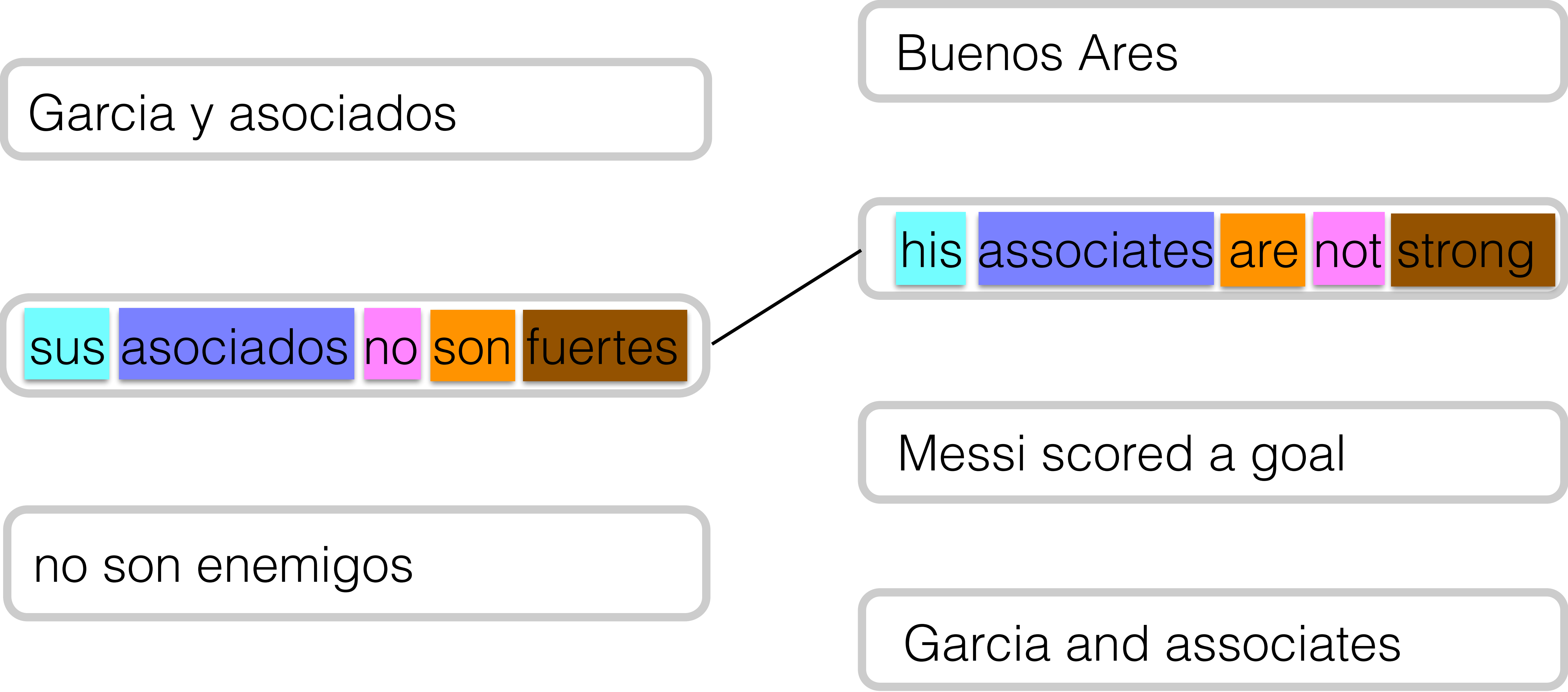
no son enemigos

Garcia and associates

iteration 1

Idea

- “**find one, get more**” (Fung and Cheung, 2004) **iteratively**



Buenos Ares

Garcia y asociados

his associates are not strong

sus asociados no son fuertes

Messi scored a goal

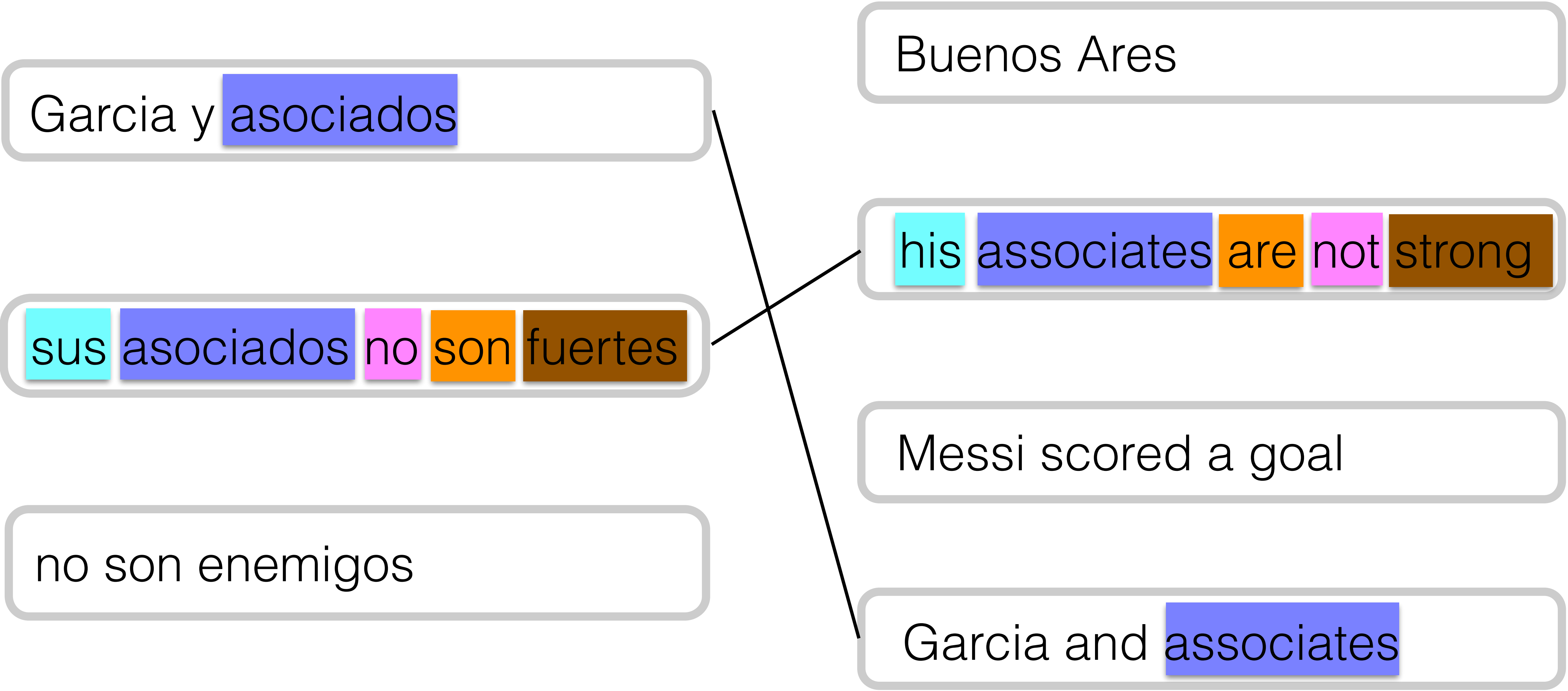
no son enemigos

Garcia and associates

iteration 2

Idea

- "**find one, get more**" (Fung and Cheung, 2004) **iteratively**



Garcia y asociados

sus asociados no son fuertes

no son enemigos

Buenos Ares

his associates are not strong

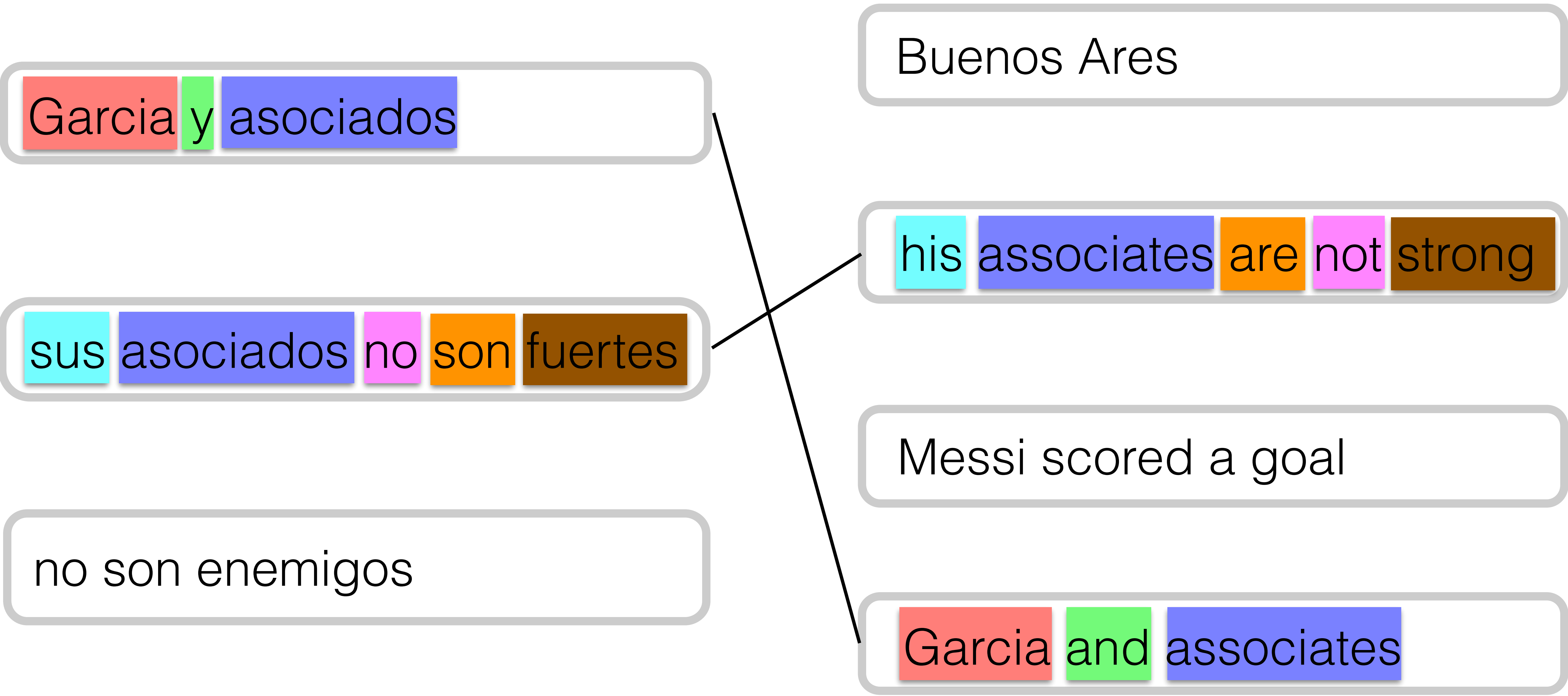
Messi scored a goal

Garcia and associates

iteration 3

Idea

- "**find one, get more**" (Fung and Cheung, 2004) **iteratively**



Garcia y asociados

sus asociados no son fuertes

no son enemigos

Buenos Ares

his associates are not strong

Messi scored a goal

Garcia and associates

iteration 4

Model

- Extending IBM models in a non-parallel scenario

phrase matching

$$P(F|E;\boldsymbol{\theta}) = \sum_{\mathbf{m}} \frac{p(T|S)}{(S+1)^T} \prod_{t=1}^{T} P(\mathbf{f}^{(t)}|\mathbf{e}^{(\mathbf{m}_t)};\boldsymbol{\theta})$$

word alignment

$$\begin{aligned} & P(\mathbf{f}^{(t)}|\mathbf{e}^{(\mathbf{m}_t)};\boldsymbol{\theta}) \\ = \quad & \delta(\mathbf{m}_t, 0)\epsilon + \\ & (1-\delta(\mathbf{m}_t,0)) \frac{p(J^{(t)}|I^{(\mathbf{m}_t)})}{(I^{(\mathbf{m}_t)}+1)^{J^{(t)}}} \prod_{j=1}^{J^{(t)}} \sum_{i=0}^{I^{(\mathbf{m}_t)}} p(\mathbf{f}_j^{(t)}|\mathbf{e}_i^{(\mathbf{m}_t)}) \end{aligned}$$

Training Objective

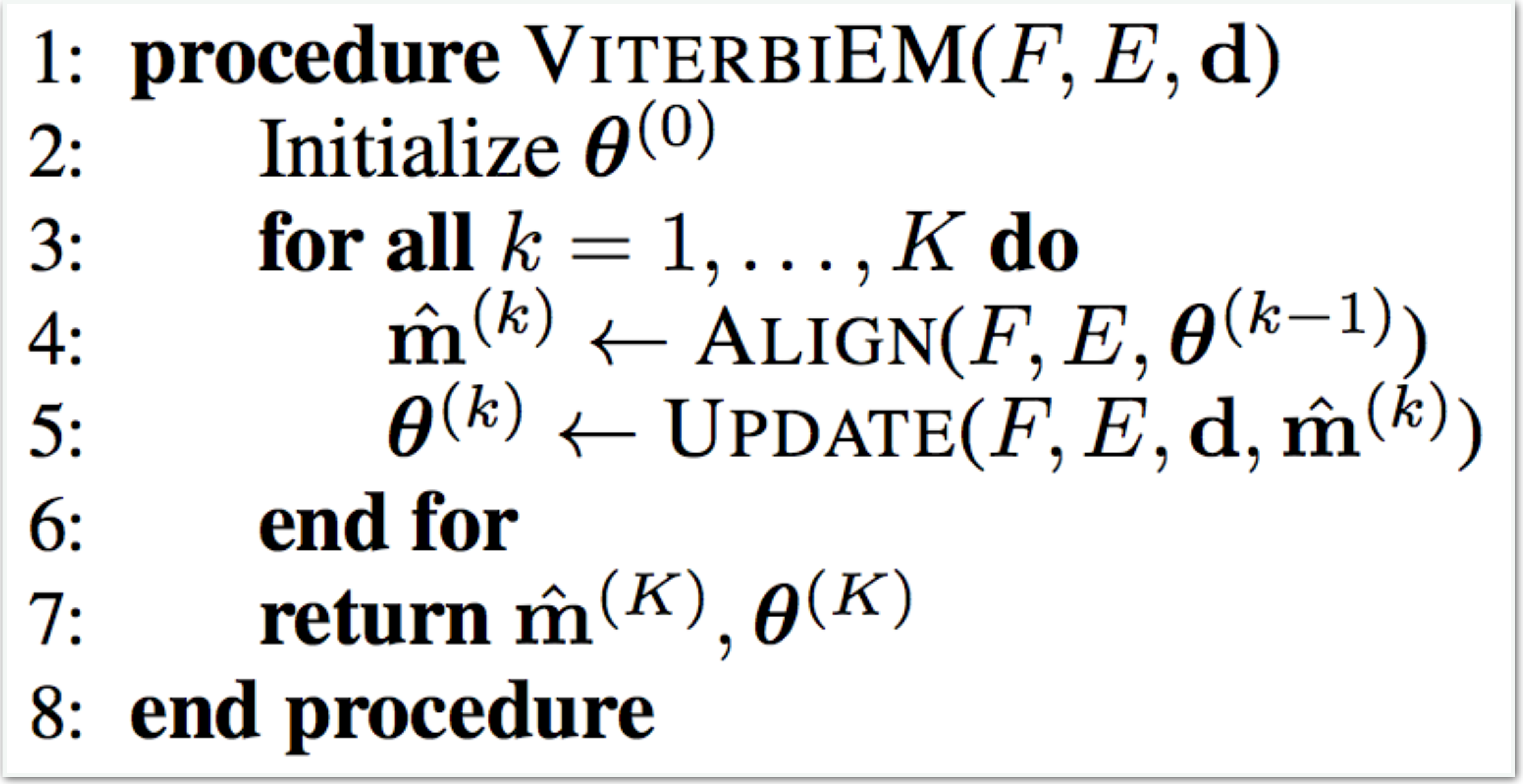
- Objective: maximizing the likelihood of training data
- We use a seed dictionary to inject prior knowledge

$$\begin{aligned} J(\boldsymbol{\theta}) \quad = \quad & P(F|E;\boldsymbol{\theta}) - \\ & \sum_{f}\sum_{e}\sigma(f,e,\mathbf{d})\log\frac{\sigma(f,e,\mathbf{d})}{p(f|e)} - \\ & -\sum_{I}\lambda_I\Big(\sum_{J}p(J|I)-1\Big) - \\ & -\sum_{e}\gamma_e\Big(\sum_{f}p(f|e)-1\Big) \end{aligned}$$

seed dictionary

Training Algorithm

- It is intractable to calculate expected counts exactly
- We resort to a Viterbi EM algorithm



```
procedure ViterbiEM($F, E, \mathbf{d}$)
    Initialize $\boldsymbol{\theta}^{(0)}$
    for all $k = 1, \ldots, K$ do
        $\hat{\mathbf{m}}^{(k)} \leftarrow$ Align($F, E, \boldsymbol{\theta}^{(k-1)}$)
        $\boldsymbol{\theta}^{(k)} \leftarrow$ Update($F, E, \mathbf{d}, \hat{\mathbf{m}}^{(k)}$)
    end for
    return $\hat{\mathbf{m}}^{(K)}, \boldsymbol{\theta}^{(K)}$
end procedure
```

Calculating Viterbi Matching

- Viterbi matching can be calculated independently
- We use information retrieval techniques for speed-up

$$\hat{\mathbf{m}}_t = \begin{cases} \tilde{\mathbf{m}}_t & \text{if } P(\mathbf{f}^{(t)} | \mathbf{e}^{(\tilde{\mathbf{m}}_t)}; \boldsymbol{\theta}) > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

$$\tilde{\mathbf{m}}_t = \operatorname*{argmax}_{s \in \{1,\ldots,S\}} \left\{ \frac{p(J^{(t)} | I^{(s)})}{(I^{(s)} + 1)^{J^{(t)}}} \prod_{j=1}^{J^{(t)}} \sum_{i=0}^{I^{(s)}} p(\mathbf{f}_j^{(t)} | \mathbf{e}_i^{(s)}) \right\}$$

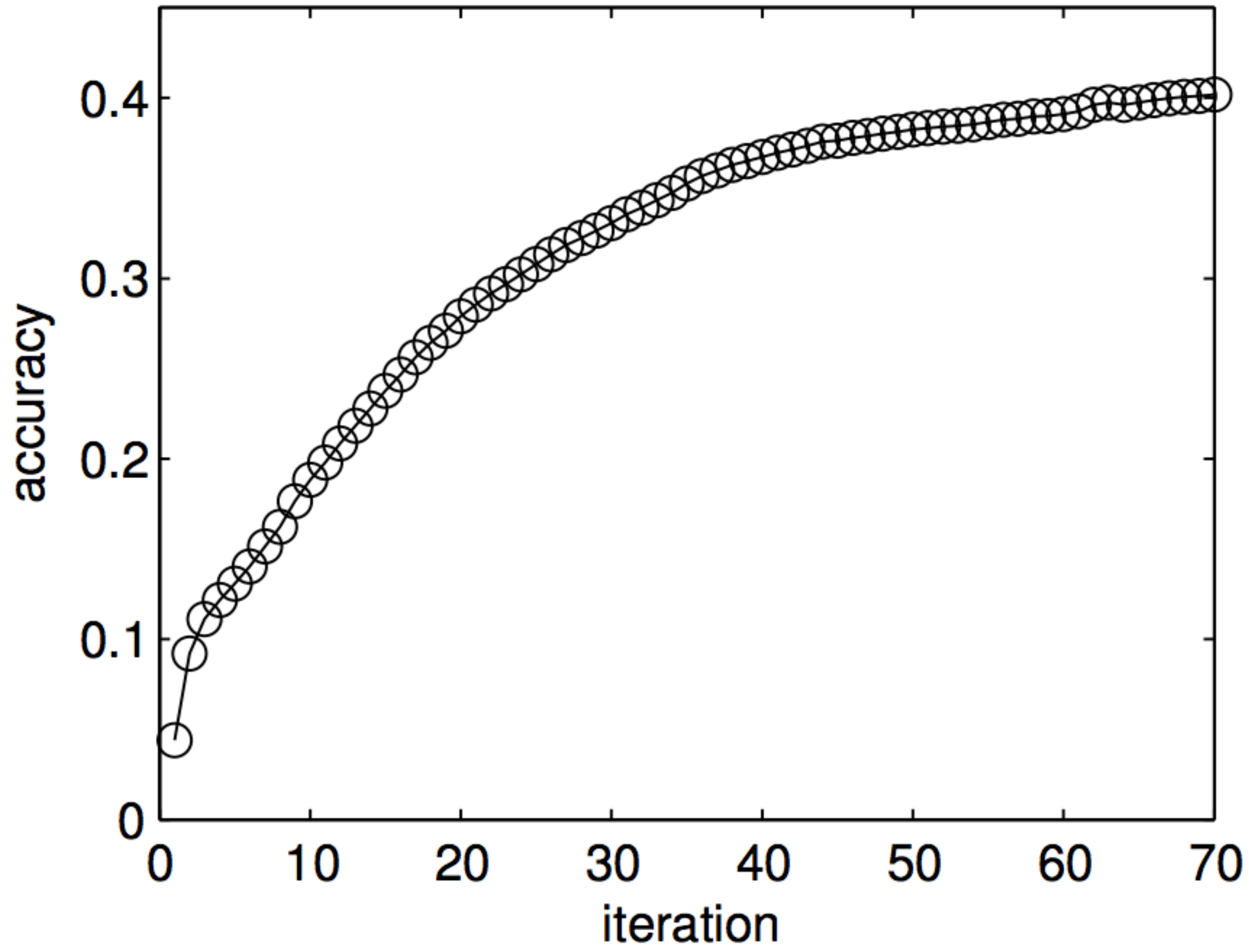
Updating Model Parameters

- Model parameters are updated based on the Viterbi matching

parameter	expected count
$p(J\|I)$	$\sum_{t=1}^{T}(1-\delta(\hat{\mathbf{m}}_t,0))\delta(J^{(t)},J)\delta(I^{(\hat{\mathbf{m}}_t)},I)$
$p(f\|e)$	$\sum_{t=1}^{T}(1-\delta(\hat{\mathbf{m}}_t,0))\frac{p(f\|e)}{\sum_{i=0}^{I^{(\hat{\mathbf{m}}_t)}} p(f\|\mathbf{e}_i^{(\hat{\mathbf{m}}_t)})}$ $\times\sum_{j=1}^{J^{(t)}}\delta(f,\mathbf{f}_j^{(t)})\sum_{i=0}^{I^{(\hat{\mathbf{m}}_t)}}\delta(e,\mathbf{e}_i^{(\hat{\mathbf{m}}_t)})$ $+\sigma(f,e,\mathbf{d})$

Matching Evaluation

- Non-parallel corpora: 20K Chinese + 40K English phrases
- seed dictionary: 1K entries



Translation Evaluation

- Using learned parallel phrases to train translation models

iteration	corpus	lexicon	BLEU
0	7.4K	4.1K	7.68
1	43.4K	6.6K	11.21
2	102.4K	9.3K	12.65
3	135.3K	10.5K	12.77
4	148.2K	11.0K	13.23
5	153.7K	11.3K	13.40

Examples

Chinese	其 主要 产品 是 食物
English	its main products are food
Chinese	美国 既定 的 战略 计划
English	the set US strategic plan
Chinese	全球化 的 趋势
English	the trends of globalization
Chinese	强化 联合 作战 功能
English	strengthening joint combat functions
Chinese	伊斯兰教 ， 天主教 和 基督教
English	Islam , Catholicism , and Christianity

Conclusion and Future Work

- **Conclusion**
    - We have presented a framework for learning parallel lexicons and phrases from non-parallel corpora
    - Useful for machine translation and bilingual lexicography
- **Future Work**
    - Extend to more sophisticated alignment models
    - Scale to large-scale data

Thanks