Introduction of MSRA Named Entity Corpus

1 Named Entity (NE) Tag-Set

Five NE categories with total 30 subcategories have been designed in the MSRA NE tag-set (see Table 1 below). The detail definition of each subcategory could be found in the spec, Tokenization Guidelines of Chinese Text.

Category	Subcategory	Tag-set of Format-1	Tag-set of Format-2
NAMEX	Person	P	PERSON
	Location	L	LOCATION
	Organization	O	ORGANIZATION
TIMEX	Date	dat	DATE
	Duration	dur	DURATION
	Time	tim	TIME
NUMEX	Percent	per	PERCENT
	Money	mon	MONEY
	Frequency	fre	FREQUENCY
	Integer	int	INTEGER
	Fraction	fra	FRACTION
	Decimal	dec	DECIMAL
	Ordinal	ord	ORDINAL
	Rate	rat	RATE
MEASUREX	Age	age	AGE
	Weight	wei	WEIGHT
	Length	len	LENGTH
	Temperature	tem	TEMPERATURE
	Angle	ang	ANGLE
	Area	are	AREA
	Capacity	cap	CAPACITY
	Speed	spe	SPEED
	Acceleration	acc	ACCELERATION
	Other measures	mea	MEASURE
ADDREX	Email	ema	EMAIL
	Phone	pho	PHONE
	Fax	fax	FAX
	Telex	tel	TELEX
	WWW	www	WWW
	Postalcode	pos	POSTALCODE

Table 1 Tag-set of Named Entities

2 Annotated Examples

While the format-1 is annotator oriented, the format-2 is XML-based which is the annotated format of the training and test corpora for the 3rd SAGHAN Bakeoff. Following are some annotated examples in these two formats respectively.

[Example-1]

Format-1:

[dat 6月29日] 、 [dat 30日] [tim 晚上] ， [L 北京市] 下 了 [int 两场] 大雨 ， 笔者 居住 的 宿舍 楼 前 ， 宽 [len 六七米] 、 长 [len 30多米] 的 路 上 ， 积水 达 膝盖 之上 。

Format-2:

<sentence>

<w><TIMEX TYPE="DATE"> 6 月 2 9 日 </TIMEX></w><w> 、 </w><w><TIMEX TYPE="DATE"> 3 0 日 </TIMEX></w><w><TIMEX TYPE="TIME"> 晚 上 </TIMEX></w><w>，</w><w><NAMEX TYPE="LOCATION">北京市</NAMEX></w><w> 下 </w><w> 了 </w><w><NUMEX TYPE="INTEGER"> 两 场 </NUMEX></w><w> 大 雨 </w><w>，</w><w>笔者</w><w>居住</w><w>的</w><w>宿舍</w><w>楼</w><w>前 </w><w> ， </w><w> 宽 </w><w><MEASUREX TYPE="LENGTH"> 六 七 米 </MEASUREX></w><w>、</w><w>长</w><w><MEASUREX TYPE="LENGTH"> 3 0 多米 </MEASUREX></w><w>的</w><w>路</w><w>上</w><w>，</w><w>积水</w><w>达 </w><w>膝盖</w><w>之上</w><w>。</w>

</sentence>

[Example-2]

[dat 6月中下旬] ， 笔者 到 [L 意大利] 、 [L 西班牙] 等 国 访问 时 ， 一个 很 深 的 感受 是 [L 意] 、[L 西] 两国 的 高速公路 非常 发达 ， 东西南北 ， 纵横 成 网 ， 四通八达 。

<sentence>

<w><TIMEX TYPE="DATE"> 6 月中下旬</TIMEX></w><w>，</w><w>笔者</w><w>到 </w><w><NAMEX TYPE="LOCATION">意大利</NAMEX></w><w>、</w><w><NAMEX TYPE="LOCATION">西班牙</NAMEX></w><w>等</w><w>国</w><w>访问</w><w>时 </w><w>， </w><w> 一个 </w><w> 很 </w><w> 深 </w><w> 的 </w><w> 感 受 </w><w> 是 </w><w><NAMEX TYPE="LOCATION"> 意 </NAMEX></w><w> 、 </w><w><NAMEX TYPE="LOCATION">西</NAMEX></w><w>两国</w><w>的</w><w>高速公路</w><w>非常</w><w>发达</w><w>，</w><w>东西南北</w><w>，</w><w>纵横</w><w>成</w><w>网</w><w>，</w><w>四通八达</w><w>。</w>

</sentence>

[Example-3]

近 [dur 四年] 来 ， 公司 [int 两个] 文明 建设 >成果 显著 ， 主要 经济 指标 保持 以 [per 3 0 %] 的 速度 连年 递增 ， 先后 评为 全国 [int 五 百家] 最佳 经济效益 企业 、 [L 北京市] 经济 [int 百] 强 企业 、 首都 文 明 单位 标兵 。

<sentence>

<w> 近 </w><w><TIMEX TYPE="DURATION"> 四 年 </TIMEX></w><w> 来 </w><w> ，</w><w> 公 司 </w><w><NUMEX TYPE="INTEGER"> 两 个 </NUMEX></w><w> 文 明 </w><w>建设</w><w>成果</w><w>显著</w><w>，</w><w>主要</w><w>经济</w><w>指 标 </w><w> 保 持 </w><w> 以 </w><w><NUMEX TYPE="PERCENT"> 3 0 % </NUMEX></w><w>的</w><w>速度</w><w>连年</w><w>递增</w><w>，</w><w>先后 </w><w> 被 </w><w> 评 为 </w><w> 全 国 </w><w><NUMEX TYPE="INTEGER"> 五 百 家 </NUMEX></w><w> 最佳 </w><w>经济效益</w><w> 企业 </w><w>、</w><w><NAMEX TYPE="LOCATION"> 北 京 市 </NAMEX></w><w> 经 济 </w><w><NUMEX TYPE="INTEGER"> 百 </NUMEX></w><w> 强 </w><w> 企 业 </w><w> 、 </w><w> 首 都 </w><w>文明</w><w>单位</w><w>标兵</w><w>。</w>

</sentence>

[Example-4]

[P 邓小平] 同志 [fre 多次] 强调 ， 搞 现代化 建设 、 经济 建设 必须 有 政治 保证 。

<sentence>

<w><NAMEX TYPE="PERSON"> 邓 小 平 </NAMEX></w><w> 同 志 </w><w><NUMEX TYPE="FREQUENCY">多次</NUMEX></w><w>强调</w><w>，</w><w>搞</w><w>现代化</w><w>建设</w><w>、</w><w>经济</w><w>建设</w><w>必须</w><w>有</w><w>政治</w><w>保证</w><w>。</w>

</sentence>

[Example-5]

我们 [O 北京城建集团一公司] 是 [dat 1 9 8 3 年] 由 基建 工程兵 集体 转业 组建 的 。

<sentence>

<w> 我 们 </w><w><NAMEX TYPE="ORGANIZATION"> 北 京 城 建 集 团 一 公 司 </NAMEX></w><w>是</w><w><TIMEX TYPE="DATE">1 9 8 3 年</TIMEX></w><w>由 </w><w> 基 建 </w><w> 工 程 兵 </w><w> 集 体 </w><w> 转 业 </w><w> 组 建 </w><w> 的 </w><w>。</w>

</sentence>

[Example-6]

领导干部 喜欢 喝酒 ， 他 手下 有 [int 几个] 『 [P 刘伶] 』 、『 [P 李 白] 』 ， 也 就 确定 无疑 了 。

<sentence>

<w>领导干部</w><w>喜欢</w><w>喝酒</w><w>，</w><w>他</w><w>手下</w><w>有

</w><w><NUMEX TYPE="INTEGER"> 几 个 </NUMEX></w><w> 『 </w><w><NAMEX TYPE="PERSON"> 刘伶 </NAMEX></w><w> 』 </w><w> 、 </w><w> 『 </w><w><NAMEX TYPE="PERSON"> 李白 </NAMEX></w><w> 』 </w><w> ， </w><w> 也 </w><w> 就 </w><w> 确定 </w><w> 无疑 </w><w> 了 </w><w> 。 </w>
</sentence>

[Example-7]
只有 这样 ， 才 能 顺乎 潮流 ， 经受 挑战 ， 实现 [L 中华]民族 的 伟大 复兴 。
<sentence>
<w> 只有 </w><w> 这样 </w><w> ， </w><w> 才 </w><w> 能 </w><w> 顺乎 </w><w> 潮流 </w><w> ， </w><w> 经受 </w><w> 挑战 </w><w> ， </w><w> 实现 </w><w><NAMEX TYPE="LOCATION"> 中华 </NAMEX> 民族 </w><w> 的 </w><w> 伟大 </w><w> 复兴 </w><w> 。 </w>
</sentence>

[Example-8]
科学研究 证明 ， 人类 在 最近 [dur 三十年] 所 获得 的 知识 约 等于 过去 [dur 两千年] 之 总和 ， 而 未来 若干 年 内 科技 和 知识 还 会 在 许多 领域 出现 更为 惊人 的 突破 。
<sentence>
<w> 科学研究 </w><w> 证明 </w><w> ， </w><w> 人类 </w><w> 在 </w><w> 最近 </w><w><TIMEX TYPE="DURATION"> 三十年 </TIMEX></w><w> 所 </w><w> 获得 </w><w> 的 </w><w> 知识 </w><w> 约 </w><w> 等于 </w><w> 过去 </w><w><TIMEX TYPE="DURATION"> 两千年 </TIMEX></w><w> 之 </w><w> 总和 </w><w> ， </w><w> 而 </w><w> 未来 </w><w> 若干 </w><w> 年 </w><w> 内 </w><w> 科技 </w><w> 和 </w><w> 知识 </w><w> 还 </w><w> 会 </w><w> 在 </w><w> 许多 </w><w> 领域 </w><w> 出现 </w><w> 更为 </w><w> 惊人 </w><w> 的 </w><w> 突破 </w><w> 。 </w>
</sentence>

[Example-9]
[L 建行] 个人 住房 贷款 [mon 万元] 利息 对比 表
<sentence>
<w><NAMEX TYPE="ORGANIZATION"> 建行 </NAMEX></w><w> 个人 </w><w> 住房 </w><w> 贷款 </w><w><NUMEX TYPE="MONEY"> 万元 </NUMEX></w><w> 利息 </w><w> 负担 </w><w> 对比 </w><w> 表 </w>
</sentence>

[Example-10]
在 本届 “ [L 灵山] 杯 ” 评选 中 ， 有 [int 二十] 人 获 一等奖 ， [int 三十四] 人 获 [ord 二等] 奖 ， [ord 三等] 奖 [int 二十七名] 。
<sentence>
<w> 在 </w><w> 本 届 </w><w> “ </w><w><NAMEX TYPE="LOCATION"> 灵山 </NAMEX></w><w> 杯 </w><w> ” </w><w> 评选 </w><w> 中 </w><w> ， </w><w> 有

</w><w><NUMEX TYPE="INTEGER">二十</NUMEX></w><w>人</w><w>获</w><w>一等奖</w><w>，</w><w><NUMEX TYPE="INTEGER">三十四</NUMEX></w><w>人</w><w>获</w><w><NUMEX TYPE="ORDINAL">二等</NUMEX></w><w>奖</w><w>，</w><w><NUMEX TYPE="ORDINAL">三等</NUMEX></w><w>奖</w><w><NUMEX TYPE="INTEGER">二十七名</NUMEX></w><w>。</w>
</sentence>

[Example-11]
[dur 二十年] 来 ， 以 经济 建设 为 中心 的 思想 比较 牢固 ， 这是 一个 很 大 的 进步 。
<sentence>
<w><TIMEX TYPE="DURATION">二十年</TIMEX></w><w>来</w><w>，</w><w>以</w><w>经济</w><w>建设</w><w>为</w><w>中心</w><w>的</w><w>思想</w><w>比较</w><w>牢固</w><w>，</w><w>这是</w><w>一个</w><w>很</w><w>大</w><w>的</w><w>进步</w><w>。</w>
</sentence>

[Example-12]
[dat 7月1日] ， [O 中国人民银行] [ord 第五次] 降低 金融 机构 存 贷款 利率 ， 中长期 贷款 利率 降幅 达到 或 超过 [per 2个百分点] ， 将 有助于 刺激 投资 需求 ， 加快 我国 基础 设施 建设 。
<sentence>
<w><TIMEX TYPE="DATE">7 月 1 日</TIMEX></w><w>，</w><w><NAMEX TYPE="ORGANIZATION">中 国 人 民 银 行</NAMEX></w><w><NUMEX TYPE="ORDINAL">第五次</NUMEX></w><w>降低</w><w>金融</w><w>机构</w><w>存</w><w>贷款</w><w>利率</w><w>，</w><w>中长期</w><w>贷款</w><w>利率</w><w> 降 幅 </w><w> 达 到 </w><w> 或 </w><w> 超 过 </w><w><NUMEX TYPE="PERCENT">2个百分点</NUMEX></w><w>，</w><w>将</w><w>有助于</w><w>刺激</w><w>投资</w><w>需求</w><w>，</w><w>加快</w><w>我国</w><w>基础</w><w>设施</w><w>建设</w><w>。</w>
</sentence>

[Example-13]
[O 北京大学光华管理学院] 博士生 [P >张后奇]
<sentence>
<w><NAMEX TYPE="ORGANIZATION">北京大学光华管理学院</NAMEX></w><w>博士生</w><w><NAMEX TYPE="PERSON">张后奇</w>
</sentence>

[Example-14]
[L 中国建设银行房贷部] 总经理 [P 李庆振]
<sentence>
<w><NAMEX TYPE="ORGANIZATION">中国建设银行房贷部</NAMEX></w><w>总经理</w><w><NAMEX TYPE="PERSON">李庆振</w>

</sentence>

[Example-15]

我们 一定 要 认真 做好 [dat 下半年] 的 工作 ， 确保 [O 首钢总公司 全年 实现 利润 [mon 9．4亿元] 的 目标 ， 并 >力争 多 完成 一些 。

<sentence>

<w> 我 们 </w><w> 一 定 </w><w> 要 </w><w> 认 真 </w><w> 做 好 </w><w><TIMEX TYPE="DATE"> 下 半 年 </TIMEX></w><w> 的 </w><w> 工 作 </w><w> ， </w><w> 确 保 </w><w><NAMEX TYPE="ORGANIZATION"> 首 钢 总 公 司 </NAMEX></w><w> 全 年 </w><w> 实 现 </w><w> 利 润 </w><w><NUMEX TYPE="MONEY"> 9 ． 4 亿 元 </NUMEX></w><w> 的 </w><w> 目 标 </w><w> ， </w><w> 并 </w><w> 力 争 </w><w> 多 </w><w>完成</w><w>一些</w><w>。</w>

</sentence>

[Example-16]

[O 县委] 决定 选派 任 了 [dur 八年] [O 城建局] 长 的 [P 周欣光] 担任 [老干部局] 长 。

<sentence>

<w><NAMEX TYPE="ORGANIZATION"> 县 委 </NAMEX></w><w> 决 定 </w><w> 选 派 </w><w> 任 </w><w> 了 </w><w><TIMEX TYPE="DURATION"> 八 年 </TIMEX></w><w><NAMEX TYPE="ORGANIZATION"> 城 建 局 </NAMEX></w><w> 长 </w><w> 的 </w><w><NAMEX TYPE="PERSON"> 周 欣 光 </NAMEX></w><w> 担 任 </w><w><NAMEX TYPE="ORGANIZATION">老干部局</NAMEX></w><w>长</w><w>。</w>

</sentence>

[Example-17]

[L 喇嘛寺村]/ 地处 [L 承德避暑山庄] ， [L 山庄] 寺庙 林立 ， 僧侣 穿梭 ， [L 山庄] [L 外八庙] 的 [ord 第一个] 庙 就 是 [L 喇嘛寺] 。

<sentence>

<w><NAMEX TYPE="LOCATION">喇嘛寺村</NAMEX></w><w>地处</w><w><NAMEX TYPE="LOCATION"> 承 德 避 暑 山 庄 </NAMEX></w><w> ， </w><w><NAMEX TYPE="LOCATION">山庄</NAMEX></w><w>寺庙</w><w>林立</w><w>，</w><w>僧侣 </w><w> 穿 梭 </w><w> ， </w><w><NAMEX TYPE="LOCATION"> 山 庄 </NAMEX></w><w><NAMEX TYPE="LOCATION"> 外 八 庙 </NAMEX></w><w> 的 </w><w><NUMEX TYPE="ORDINAL">第一个</NUMEX></w><w>庙</w><w>就</w><w>是</w><w><NAMEX TYPE="LOCATION">喇嘛寺</NAMEX></w><w>。</w>

</sentence>

3 Statistics of Training Corpus

# of sentences	# of words (including punctuations)	# of NE tags	Ave. # of words per sentence	Ave. # of NE tags per sentence
46,364	1,265,764	118,643	27.30	2.56

Table 2 Size of MSRA Training Corpus

Number of L	Number of O	Number of P	Total number
36,860	20,584	17,615	75,059

Table 3 Number of NAMEX Tags

Number of dat	Number of dur	Number of tim	Total number
12,661	3,928	1,335	17,924

Table 4 Number of TIMEX Tags

# of int	# of ord	# of mon	# of per	# of fre
10,433	3,506	2,346	1,786	939
# of rat	# of fra	# of dec	Total number	
622	467	340	20,439	

Table 5 Number of NUMEX Tags

# of tem	# of len	# of age	# of are	# of wei	Total number
2,741	735	612	439	424	5,165
# of mea	# of cap	# of ang	# of spe	# of acc	
137	46	18	13	0	

Table 6 Number of MEASUREX Tags

# of pho	# of www	# of pos	# of fax	# of tel	Total number
40	13	3	0	0	56

Table 7 Number of ADDREX Tags

<END>